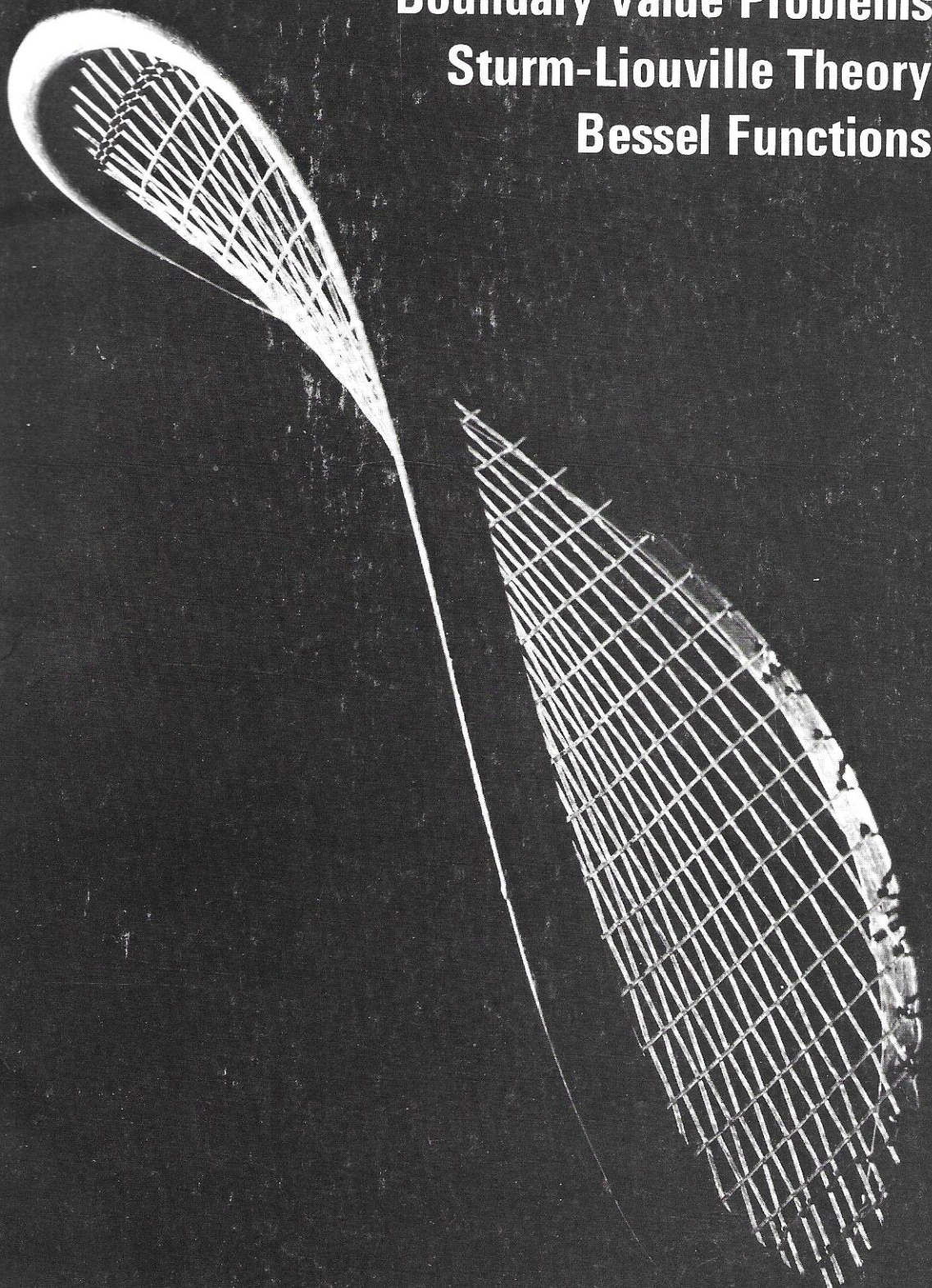
M321 11, 13 and 14

THE OPEN UNIVERSITY

Mathematics : A Third Level Course

Partial Differential Equations of Applied Mathematics Units 11, 13 and 14

Boundary Value Problems
Sturm-Liouville Theory
Bessel Functions



THE OPEN UNIVERSITY

Mathematics: A Third Level Course

Partial Differential Equations of Applied Mathematics
Units 11, 13 and 14

BOUNDARY VALUE PROBLEMS
STURM-LIOUVILLE THEORY
BESSEL FUNCTIONS

Prepared by the Course Team

The Open University Press

Unit 11 Finite Difference Methods III: Boundary Value Problems

The Open University Press, Walton Hall, Milton Keynes.

First published 1974

Produced in Great Britain by
Technical Filmsetters Europe Limited, 76 Great Bridgewater Street, Manchester M1 5JY

ISBN 0 335 01253 1

This text forms part of the correspondence element of an Open University Third Level Course. The complete list of units in the course is given at the end of this text.

For general availability of supporting material referred to in this text, please write to the Director of Marketing, The Open University, P.O. Box 81, Milton Keynes, MK7 6AT.

Further information on Open University courses may be obtained from The Admissions Office, The Open University, P.O. Box 48 Milton Keynes, MK7 6AB.

1.1

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differential Equations* (Oxford, 1971).
H. F. Weinberger, *A First Course in Partial Differential Equations* (Xerox, 1965).

It is essential to have these books; the course is based on them and will not make sense without them. They are referred to in the text as *S* and *W* respectively.

Unit 11 is based on *S*: Chapter 5, pages 131 to 137, 143 to 149.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics.* References to Open University courses in mathematics take the form:

Unit M100 13, Integration II for the Mathematics Foundation Course.

Unit M201 23, The Wave Equation for the Linear Mathematics Course.

11.0 INTRODUCTION

In *Unit 5, Initial Value Problems* and *Unit 8, Stability* we dealt with the finite-difference method for solving initial value problems in one space dimension. We now turn our attention to pure boundary value problems in two space dimensions. Here we seek the solution of a partial differential equation in some finite region of the xy-plane bounded by a closed curve on which a (boundary) condition is specified at every point. A typical problem on which we shall concentrate throughout this unit is *Poisson's equation*,

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f,$$

in some rectangular region.

We begin by applying the finite-difference techniques of *Unit 5* to two problems which illustrate that there are two distinct topics for investigation in the numerical solution of boundary value problems. The first and most obvious task is to search for good numerical methods for solving the finite-difference equations which arise. The second task is familiar from our work in *Unit 8*, namely the investigation of the accuracy of the finite-difference solution as a solution to the boundary value problem. This latter investigation will, as always, raise the questions of stability and convergence. In Section 11.2 we shall look at these questions in some detail for Poisson's equation, where we shall find that a *maximum principle* applies just as in the analysis of boundary value problems for differential equations in *Unit 3, Elliptic and Parabolic Equations*. The maximum principle will enable us to find some bounds on the error in our numerical solution.

The remainder of the unit deals with the numerical solution of the finite-difference equations. Our two examples highlight the important fact that a finite-difference method applied to boundary value problems gives rise to very large sets of simultaneous equations which are linear if the differential equation is linear. The associated matrices are of a special *sparse* form, which means that they contain a significant majority of zero elements. We know from *Unit M201 8, Numerical Solution of Simultaneous Algebraic Equations*, that linear equations can be solved by various *direct* methods, such as Gauss elimination. With such methods it is sometimes possible to take sparsity into account, as for example in *Unit 5* where we modified the elimination method to give a recurrence relation for solving tridiagonal systems of equations. Unfortunately the sparse matrices associated with elliptic partial differential equations have a more complicated structure, and it is difficult to find economic direct methods of solution.

There is, however, another important class of methods for solving linear equations, namely *iterative* methods which we first considered in *Unit M100 28, Linear Algebra IV*. It turns out that iterative methods are specially suitable for our particular class of large sparse matrices, and we shall look at iterative processes in general terms as well as discussing specific methods.

11.1 TWO EXAMPLES

The first reading passage begins with a review of some general properties of boundary value problems which were discussed in detail in *W:* Sections 10 and 11.

READ S: page 131, **Introduction** *to page 137, line 6.*

Notes

(i) *S: page 132, line 9*
A specified $\partial\phi/\partial n$ on C does not determine ϕ uniquely in S. If ϕ is a solution, then $\phi + A$ is a solution for any constant A (see *W: page 54*).

(ii) *S: page 133, line 5*
Although we are now dealing with pure boundary value problems we still approximate the differential equation by finite-difference schemes similar to those for initial value problems derived in *Unit 5*. Employing the usual central-difference replacement for second-order derivatives we can write the required finite-difference scheme as

$$\frac{1}{h^2}(\delta_x^2 + \delta_y^2)\phi + 2 = 0,$$

where we use the same mesh spacing h in both the x- and y-directions.

(iii) *S: page 134, line 1*
In order to improve the accuracy we could increase the number of mesh points (thereby reducing h). This increases the number of simultaneous equations which have to be solved.

(iv) *S: page 134, line 3*
It can be shown that the solution of Poisson's equation near a re-entrant corner (where the interior angle exceeds 180°) is oscillatory, making our finite-difference methods inaccurate there. (See L. Fox, ed., *Numerical Solution of Ordinary and Partial Differential Equations*, p. 303.)

(v) *S: page 134, lines 12 to 17*
You need not worry about the derivation of the equation on line 15, but note that putting $y = \lambda^{\frac{1}{2}}\bar{y}$ (provided λ is a positive constant) yields

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial \bar{y}^2} = -\frac{q}{K},$$

which is Poisson's equation.

(vi) *S: page 134, lines −8 and −7*
The symmetry condition yields

$$u = 0 \qquad \text{on } x = -1$$

and

$$\frac{\partial u}{\partial y} = u \qquad \text{on } y = -1.$$

It also implies that there will be no heat flow across Oy and Ox; hence,

$$\frac{\partial u}{\partial x} = 0 \qquad \text{on } Oy$$

and

$$\frac{\partial u}{\partial y} = 0 \qquad \text{on } Ox.$$

(vii) *S: page 136, lines −2 and −1*
The matrix **A** is an example of a type of matrix which occurs frequently in this subject. In its partitioned form we see that **A** has a tridiagonal appearance and

we therefore refer to it as a **block tridiagonal matrix**. Note also that each of the diagonal blocks in **A** is a tridiagonal matrix.

SAQ 1

If, in Equation (5.2) on *S: page 134*, we employ the transformation $y = \sqrt{3}\,\bar{y}$ to obtain Poisson's equation (note (v) above), what changes must be made to the boundary values to retain the same problem? What differences will there be in the numerical solution of this new problem compared to the solution given in *S: pages 134 to 137*, if the same mesh is used?

(Solution on p. 35.)

SAQ 2

S: page 161, Exercise 1 (Omit the last part of the question which is on *page 162*.)

(Solution on p. 35.)

SAQ 3

S: page 162, Exercise 2
Note that the mesh point numbering given in the question is different from that in Fig. 5.2; the reference to Fig. 5.2 is for illustration only.

(Solution on p. 36.)

11.2 THE FIVE-POINT FORMULA

11.2.0 Introduction

The two examples of the previous section have illustrated two important facts associated with the solution of boundary value problems by finite-difference methods. The first is that the boundary value problem is approximated by an algebraic problem in which it is necessary to solve a system of simultaneous linear equations. The second fact is that the system of equations is generally very large. Therefore, our analysis of the numerical process divides into two distinct parts. We first need to know the conditions under which the solution of the approximating algebraic problem approaches the true solution of the differential equation; and then we need to investigate efficient methods for solving the algebraic problem. *S* deals only with solving the algebraic problem. This section, therefore, deals with the question of how well the solution of the algebraic problem approximates the true solution of the boundary value problem. There is no single technique which answers this in general. We have therefore chosen to investigate in some detail the application of one simple finite-difference scheme to the solution of Poisson's equation, which can be dealt with using a maximum principle much like the one we studied in *Unit 3*.

We look at the problem

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f \qquad (x, y) \in D, \tag{1a}$$

$$U = g \qquad (x, y) \in C, \tag{1b}$$

where D is some domain in the xy-plane and C is its boundary.

We cover D by a rectangular mesh such that the mesh spacing in the x-direction is Δx and in the y-direction is Δy. We shall approximate the Laplacian by central differences to give the **five-point formula**,

$$\left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}\right)_{i,j} \simeq \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta y)^2}.$$

Hence the finite-difference problem can be written as

$$[Lu]_{i,j} = f_{i,j} \qquad i, j \in D_\Delta, \tag{2a}$$

$$u_{i,j} = g_{i,j} \qquad i, j \in C_\Delta, \tag{2b}$$

where D_Δ is the set of mesh points interior to D, C_Δ is the set of mesh points* on C and L is the finite-difference operator given by

$$[Lu]_{i,j} = \left[\frac{\delta_x^2}{(\Delta x)^2} + \frac{\delta_y^2}{(\Delta y)^2}\right] u_{i,j}$$

$$= \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta y)^2}. \tag{3}$$

Usually we take $\Delta x = \Delta y = h$ (say) in which case the finite-difference formula reduces to

$$[Lu]_{i,j} = \frac{1}{h^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}).$$

In Section 11.2.1 we show that the solution to the finite-difference problem (2) with L defined by Equation (3) is unique. We do this by establishing a maximum principle for this problem which enables us to show that the homogeneous problem, where $f_{i,j} = g_{i,j} = 0$, has only the trivial solution $u_{i,j} = 0$. In consequence the solution to the nonhomogeneous problem is unique. The maximum principle has a further important use in that we can determine from it a bound on our approximate solution. We shall obtain this bound in Section 11.2.2, and we deduce that the chosen finite-difference scheme is convergent to the solution of the differential equation as $\Delta x, \Delta y \rightsquigarrow 0$.

* We shall assume throughout that C coincides with a union of mesh lines.

11.2.1 The Maximum Principle

In *Unit 3* (Section 3.2.1) we saw that if U is a solution of

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f \qquad (x, y) \in D$$

and $f(x, y) \geqslant 0$ in D, then if U is continuous on $D \cup C$ it attains its maximum on C. Further, in the case where $f(x, y) = 0$, which gives Laplace's equation, both the maximum and minimum values of U are attained on C. A desirable property of the solution of the finite-difference problem, if it is to be a reasonable approximation to the solution of the boundary value problem, is that it too satisfies a similar maximum principle. This is in fact the case as we show in the following theorem.

THEOREM 1

Let L be the finite-difference operator

$$\frac{\delta_x^2}{(\Delta x)^2} + \frac{\delta_y^2}{(\Delta y)^2}.$$

(a) If $v_{i,j}$ is a function defined on the set of mesh points $D_\Delta \cup C_\Delta$ which satisfies

$$Lv_{i,j} \geqslant 0 \qquad \text{for all } i, j \in D_\Delta,$$

then

$$\max_{D_\Delta} v_{i,j} \leqslant \max_{C_\Delta} v_{i,j}.$$

(b) If $v_{i,j}$ satisfies

$$Lv_{i,j} \leqslant 0 \qquad \text{for all } i, j \in D_\Delta,$$

then

$$\min_{D_\Delta} v_{i,j} \geqslant \min_{C_\Delta} v_{i,j}.$$

Proof

(a) The proof is by contradiction. Assume that at some point P_0 with coordinates $(i_0 \Delta x, j_0 \Delta y)$ we have

$$v_{i_0, j_0} = M$$

where

$$M \geqslant v_{i,j} \qquad \text{for all } i, j \in D_\Delta$$

and

$$M > v_{i,j} \qquad \text{for all } i, j \in C_\Delta.$$

That is to say, we assume that the maximum value M of the function v is attained at some point P_0 in the interior of the domain and that the boundary values are all less than M.

We can write the five-point formula as

$$Lv_0 = \frac{v_1 + v_2}{(\Delta x)^2} + \frac{v_3 + v_4}{(\Delta y)^2} - 2v_0\left[\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right] \tag{4}$$

where we have used the notation

$$v_0 = v_{i_0, j_0},$$

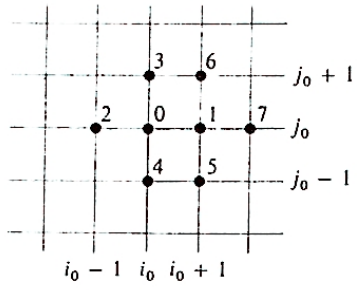
$$v_1 = v_{i_0 + 1, j_0},$$

$$v_2 = v_{i_0 - 1, j_0},$$

$$v_3 = v_{i_0, j_0 + 1},$$

$$v_4 = v_{i_0, j_0 - 1}.$$

as in the following diagram.



Now the hypothesis that $Lv_0 \geqslant 0$ gives

$$M = v_0 \leqslant \frac{1}{(\Delta x)^2 + (\Delta y)^2}\left[(\Delta y)^2 \frac{v_1 + v_2}{2} + (\Delta x)^2 \frac{v_3 + v_4}{2}\right]. \tag{5}$$

from Equation (4). But $M \geqslant v_{i,j}$ in D_Δ implies, by Equation (5), that

$$v_r = M \qquad r = 1, 2, 3, 4,$$

since the coefficients of the v_r are all positive and their sum is 1. (For if, say, $v_1 < M$ then the right-hand side of (5) is less than M.)

This means that if $v_0 \in D_\Delta$ is a maximum then so are v_1, v_2, v_3 and v_4. We now choose one of the v_r ($r = 1, 2, 3, 4$) as a new central point for the five-point formula and repeat the above argument. For example, choosing v_1 as the new central point will involve v_0, v_5, v_6 and v_7 in the five-point formula (see figure). Since we have just shown that v_1 is a maximum, so are v_5, v_6 and v_7. In this way we can include all points in D_Δ and C_Δ to show that if one point in the interior is a maximum then all points are maxima, that is

$$v_{i,j} = M \text{ at all points of } D_\Delta \text{ and } C_\Delta.$$

This contradicts the assumption that $v < M$ on C_Δ and part (a) is proved.

(b) An argument similar to (a) could be employed here but it is simpler to note that

$$\max(-v_{i,j}) = -\min v_{i,j}$$

and

$$L[-v]_{i,j} = -Lv_{i,j}.$$

Hence, if $v_{i,j}$ satisfies the hypothesis of (b) that $Lv_{i,j} \leqslant 0$ then $-v_{i,j}$ satisfies the hypothesis of (a). The conclusion of part (a) for $-v_{i,j}$ is identical to the conclusion of (b) for $v_{i,j}$.

Hence we have established the maximum and minimum principles for the five-point finite-difference replacement of the Laplacian.

If we now consider the homogeneous form of Equations (2) in which $f_{i,j} = g_{i,j} = 0$ for all $i, j \in D_\Delta \cup C_\Delta$ then Theorem 1 tells us that both the maximum and minimum values of $v_{i,j}$ are zero, and therefore the only solution is the trivial one.

This fact enables us to prove uniqueness. For, suppose that Equations (2) have two solutions u and v, and that $w \equiv u - v$. Then,

$$Lu_{i,j} = f_{i,j} \qquad i, j \in D_\Delta,$$

$$u_{i,j} = g_{i,j} \qquad i, j \in C_\Delta,$$

and

$$Lv_{i,j} = f_{i,j} \qquad i, j \in D_\Delta,$$

$$v_{i,j} = g_{i,j} \qquad i, j \in C_\Delta,$$

and therefore by subtraction

$$Lw_{i,j} = 0 \qquad i,j \in D_\Delta,$$

$$w_{i,j} = 0 \qquad i,j \in C_\Delta.$$

The above observation that the homogeneous case has only the trivial solution, $w \equiv 0$, implies that $u \equiv v$, and the solution of Equations (2) is therefore unique.

11.2.2 Error Analysis

When the finite-difference equations (2) are written down for every internal mesh point we have a set of simultaneous equations

$$A\mathbf{u} = \mathbf{b}, \tag{6}$$

where the elements of the column vector $\mathbf{u}$ are the mesh values $u_{i,j}$, A is a square matrix of order equal to the number of internal mesh points and whose coefficients depend on the finite-difference operator L, and $\mathbf{b}$ is a column vector of known elements depending on the internal mesh values of $f_{i,j}$ and the boundary values $g_{i,j}$. If $\mathbf{U}$ is the corresponding column vector of values of the *true* solution of the *differential* equation, then we define the local truncation error of our finite-difference approximation at the point i,j by the formula

$$T_{i,j} = LU_{i,j} - f_{i,j}.$$

If we have no computer storage errors (rounding errors) in the numbers $f_{i,j}$, $g_{i,j}$ and $u_{i,j}$ it is then easy to see that the vector $\mathbf{U}$ satisfies, instead of (6), the equation

$$A\mathbf{U} = \mathbf{b} + \mathbf{T} \tag{7}$$

where $\mathbf{T}$ is the column vector of local truncation errors. The column vector $\mathbf{e}$ of global errors $e_{i,j}$ where

$$e_{i,j} = U_{i,j} - u_{i,j}$$

is thus given by

$$A\mathbf{e} = \mathbf{T}, \tag{8a}$$

by subtracting Equation (6) from Equation (7). We have already shown that Equation (6) has a unique solution, so that the inverse matrix A^{-1} exists, and we can write Equation (8a) in the form

$$\mathbf{e} = A^{-1}\mathbf{T}. \tag{8b}$$

Now, just as in initial value problems, we would hope that as we steadily reduced the mesh spacings Δx and Δy our computed solutions would get closer to the true solution. Both the matrix A and (as we shall show) the vector $\mathbf{T}$ depend on Δx and Δy, and what we are hoping to show is that $\lim A^{-1}\mathbf{T} = \mathbf{0}$ (the zero vector) as $\Delta x, \Delta y \rightsquigarrow 0^+$. This will certainly happen if $\lim \mathbf{T} = \mathbf{0}$ as $\Delta x, \Delta y \rightsquigarrow 0^+$, and the inverse matrix A^{-1} is bounded as $\Delta x, \Delta y \rightsquigarrow 0^+$ (in the sense discussed in *Unit 8*, Section 8.1). The first requirement, that $\lim \mathbf{T} = \mathbf{0}$ as $\Delta x, \Delta y \rightsquigarrow 0^+$, means (as before) that our finite-difference approximations are *consistent* with the given differential equation. The requirement of a bounded A^{-1} is somewhat analogous to the concept of *stability* in our treatment of initial value problems, and if $\lim \mathbf{e} = \mathbf{0}$ as $\Delta x, \Delta y \rightsquigarrow 0^+$ we say that our method is *convergent*, again in analogy with the corresponding initial value situation.

We can obtain the local truncation error $T_{i,j}$ of our five-point finite-difference formula applied to Equation (1a) by Taylor's Theorem as explained in *Unit 5*. We find (see SAQ 4) that

$$T_{i,j} = \frac{1}{12}\left[(\Delta x)^2 \frac{\partial^4 U_{i,j}}{\partial x^4} + (\Delta y)^2 \frac{\partial^4 U_{i,j}}{\partial y^4}\right] + O((\Delta x)^4) + O((\Delta y)^4). \tag{9}$$

It follows that if the various derivatives exist throughout the domain D then $\lim T_{i,j} = 0$ as Δx and $\Delta y \rightsquigarrow 0^+$, and we have consistency.

It is also convenient to simplify expression (9) for $T_{i,j}$ and find an upper bound for it. If we use Taylor's Theorem with remainder (see *Unit 8*, note (ii) of Section 8.2.1) it is easy to show (see SAQ 4) that

$$|T_{i,j}| \leqslant \tfrac{1}{12}[M_x(\Delta x)^2 + M_y(\Delta y)^2], \tag{10}$$

where M_x and M_y are the largest values of $|\partial^4 U/\partial x^4|$ and $|\partial^4 U/\partial y^4|$ in $D \cup C$.

Let us now turn our attention to the global error. As a consequence of the following theorem, which uses the result of Theorem 1, we shall obtain a bound on the global error $e_{i,j}$.

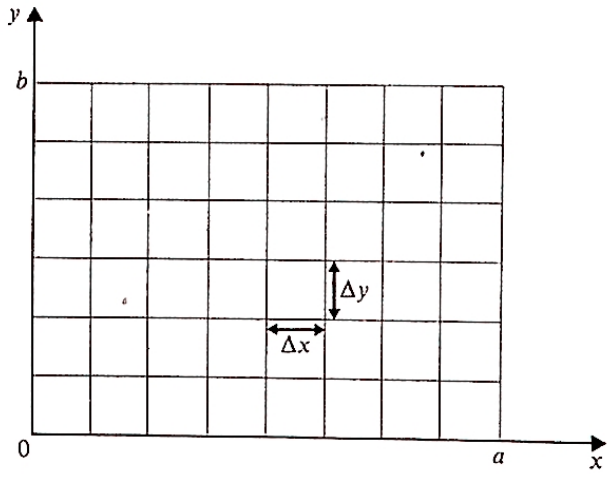
THEOREM 2

Let v be any function defined on the set of mesh points $D_\Delta \cup C_\Delta$ contained in some rectangular region $[0, a] \times [0, b]$. Then

$$\max_{D_\Delta} |v| \leqslant \max_{C_\Delta} |v| + \tfrac{1}{2}a^2 \max_{D_\Delta} |Lv|,$$

where L is the five-point finite-difference operator

$$\frac{\delta_x^2}{(\Delta x)^2} + \frac{\delta_y^2}{(\Delta y)^2}.$$



Proof

We introduce the function ϕ on $D_\Delta \cup C_\Delta$ given by

$$\phi_{i,j} = \tfrac{1}{2}(i\Delta x)^2,$$

and observe that for all $i, j \in D_\Delta \cup C_\Delta$,

$$0 \leqslant \phi_{i,j} \leqslant \tfrac{1}{2}a^2 \tag{11}$$

and

$$\begin{aligned} L\phi_{i,j} &= \tfrac{1}{2}[(i+1)^2 - 2i^2 + (i-1)^2] \\ &= 1. \end{aligned}$$

Now define two functions v^+ and v^- on $D_\Delta \cup C_\Delta$ by

$$v^\pm = \pm v + N\phi, \tag{12}$$

where

$$N = \max_{D_\Delta} |Lv_{i,j}|.$$

Clearly, operating on both sides of Equation (12) by L,

$$Lv^\pm_{i,j} = \pm Lv_{i,j} + N \geqslant 0 \qquad i, j \in D_\Delta,$$

the inequality coming from the definition of N. Applying the maximum principle (Theorem 1a) to each of $v^{\pm}$ we obtain

$$\begin{aligned} v^{\pm}_{i,j} &\leqslant \max_{C_\Delta} v^{\pm}_{i,j} \\ &\leqslant \max_{C_\Delta}[\pm v_{i,j} + N\phi_{i,j}] \qquad \text{by Equation (12)} \\ &\leqslant \max_{C_\Delta}[\pm v_{i,j}] + \tfrac{1}{2}Na^2 \qquad \text{by Equation (11).} \end{aligned}$$

The definition of $v^{\pm}_{ij}$ and the fact that $\phi_{i,j} \geqslant 0$ imply that

$$\pm v_{i,j} \leqslant v^{\pm}_{i,j}, \qquad \forall\, i,j \in D_\Delta \cup C_\Delta.$$

Hence,

$$\begin{aligned} \pm v_{i,j} &\leqslant \max_{C_\Delta}[\pm v_{i,j}] + \tfrac{1}{2}Na^2 \\ &\leqslant \max_{C_\Delta} |v| + \tfrac{1}{2}a^2N. \end{aligned}$$

Since the right-hand side of the final inequality is independent of i,j in D_Δ, the theorem follows.

[Note that we can replace $\frac{1}{2}a^2$ by $\frac{1}{2}b^2$ since the function ψ given by $\psi_{i,j} = \frac{1}{2}(j\Delta y)^2$ can be used to replace $\phi_{i,j}$ in the proof. Thus we can reduce the bound by choosing the smaller of a and b.]

The bound on the error is now easily found. The local truncation error $T_{i,j}$ is given by

$$LU_{i,j} - f_{i,j} = T_{i,j} \qquad i,j \in D_\Delta,$$

where $U_{i,j}$ is the true solution of the boundary value problem. Hence,

$$Le_{i,j} = LU_{i,j} - Lu_{i,j} = T_{i,j} \qquad i,j \in D_\Delta,$$

and also,

$$e_{i,j} = U_{i,j} - u_{i,j} = 0 \qquad i,j \in C_\Delta,$$

since the same boundary data are used in the calculation of the true solution U and the finite-difference solution u. Applying Theorem 2 to the function $e = U - u$, we obtain

$$\max_{D_\Delta}|e_{i,j}| \leqslant \tfrac{1}{2}a^2 \max_{D_\Delta}|T_{i,j}|. \tag{13}$$

The consequence of Equation (13) is that, provided the $T_{i,j}$ are finite, the $e_{i,j}$ are bounded.

Finally, our bound (10) for the truncation error shows that

$$\max|e_{i,j}| \leqslant \tfrac{1}{24}a^2[M_x(\Delta x)^2 + M_y(\Delta y)^2] \tag{14}$$

and if $\Delta x = \Delta y = h$, and M is the greater of M_x and M_y, we have the elegant result

$$\max|e_{i,j}| \leqslant \tfrac{1}{12}a^2h^2M, \tag{15}$$

showing that the *global* error is of the same order as the local truncation error. Since the right-hand side of (15) approaches zero as $h \rightsquigarrow 0$, our numerical scheme is convergent.

It is, of course, difficult to compute this error bound, since it depends on the maximum value of the fourth derivatives of the function we are trying to compute! In general we do not try to do this, but use (15) to *guarantee* convergence and to indicate the *rate* of convergence. For example, if we halve the interval h we shall reduce the error to approximately one quarter of its previous value, and this is important and useful information.

In all this analysis we have assumed that there are no computer errors in the stored values of the elements of A and $\mathbf{b}$ in the linear equation (6), that is, of the coefficients

of the operator L and the quantities $f_{i,j}$ and $g_{i,j}$. Possible uncertainties or storage errors in $f_{i,j}$ and $g_{i,j}$ can be analysed by using the maximum principle. If $\bar{u}$ is the computed solution of the problem then we can write

$$L\bar{u}_{i,j} = f_{i,j} + R'_{i,j} \qquad i,j \in D_\Delta,$$

$$\bar{u}_{i,j} = g_{i,j} + R''_{i,j} \qquad i,j \in C_\Delta,$$

where we have distinguished between the errors R' made at points in D_Δ and those R'' made in approximating the boundary data.

As before, we can obtain the expressions

$$L(U - \bar{u})_{i,j} = T_{i,j} - R'_{i,j} \qquad i,j \in D_\Delta$$

$$U_{i,j} - \bar{u}_{i,j} = -R''_{i,j} \qquad i,j \in C_\Delta$$

and applying Theorem 2 to the function $U - \bar{u}$, we obtain

$$\max_{D_\Delta} |U - \bar{u}| \leqslant \max_{C_\Delta} |R''| + \tfrac{1}{2}a^2 \max_{D_\Delta} |T - R'|,$$

and once again the error is bounded.

Alternatively we could argue that possible small uncertainties in the data will not cause significant *inherent* instability since our problem is properly posed, as we have seen in *Unit 3*.

We have also assumed that we introduce very little *induced* instability in *solving* the linear equations. We already know (*Unit M201 8*) that this is true if we use the direct method of Gauss elimination with pivoting. It can also be shown, though we shall not prove this, that no additional induced instability need be associated with our preferred methods of iteration (for our particular boundary value problems); we shall consider these methods in Section 11.3.

SAQ 4

Derive an expression for the local truncation error of the five-point formula applied to

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f$$

in a bounded domain D, and obtain a bound on its absolute value.

(Solution on p. 37.)

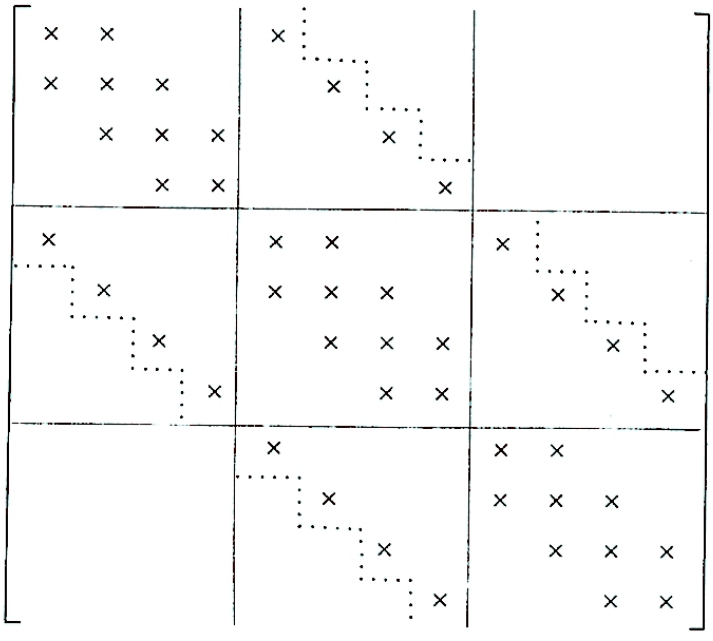
11.3 GENERAL ITERATIVE METHODS

11.3.1 Discussion

There are two classes of numerical methods for solving systems of linear algebraic equations, direct and iterative. In *direct* methods, of which Gaussian elimination with pivoting is the main example, the solution is obtained in a known number of arithmetic operations (which depends on the order of the matrix). With exact arithmetic the solution so obtained is exact, and only the unavoidable occurrence of digital computer rounding errors impairs this. In *iterative* methods, on the other hand, we obtain solutions by a process of successive approximation. We shall rarely obtain an exact solution, even with exact arithmetic, but if the method converges then we shall steadily approach the true solution, terminating the iterative sequence when our results are sufficiently accurate for our purpose.

Important considerations for iterative methods are therefore (i) to find a convergent method, and (ii) to find a method which converges reasonably quickly, i.e., one which has a reasonable *rate of convergence*. Different methods will be needed in different circumstances, and we shall rarely be able to estimate in advance the amount of computation required to achieve some specific accuracy. We shall, however, be able to perform some useful analysis for the particular types of linear equations which arise from finite-difference methods applied to our elliptic boundary value problems.

One main feature of such equations is the *sparseness* of the associated matrix, that is, the large number of zero elements in the matrix. This is manifest in the following matrix, typical of our problems, in which the nonzero elements are marked with a cross.



(1)

For such matrices direct methods have the disadvantage that they use a large amount of computer storage. Any attempt to take advantage of the many zeros tends to be frustrated by the fact that, in the process of elimination, few of the zeros within the dotted lines in the diagram can be preserved. They get "filled in" with nonzero elements, giving rise to more arithmetic than would seem to be necessary. If interchanges are necessary to avoid induced instability then even more nonzero elements are introduced in other parts of the matrix.

Iterative methods do not destroy the sparseness, and for the type of matrix shown above we can find some useful methods with satisfactory rates of convergence. In discussing these methods we shall assume that our linear equations are *well-conditioned*. They can be written in the matrix form

$$A\mathbf{u} = \mathbf{b}, \tag{2}$$

which is well-conditioned if small changes in the elements of A and $\mathbf{b}$ do not cause large changes in the solution vector $\mathbf{u}$. This is clearly a reasonable assumption if our given problem is properly posed and our finite-difference equations are such that we get nearer and nearer to the true solution as the mesh spacings are steadily reduced.

We know, moreover, from *Unit M201 8*, that the Gauss elimination method with pivoting does not suffer from induced instability, the solution obtained being the exact solution of a problem $A\mathbf{u} = \mathbf{b}$ with small perturbations in the elements of A and $\mathbf{b}$. If the problem is well-conditioned our computed result must therefore be close to the exact solution of $A\mathbf{u} = \mathbf{b}$. We can show, though we shall not prove this, that our iterative methods likewise do not suffer from induced instability.

SAQ 5

(a) Put crosses for the nonzero elements in the matrix (1) after two steps of Gauss elimination *without* interchanges.

(b) Put crosses for the nonzero elements in the matrix (1) after two steps of Gauss elimination, where the "initial" matrix has rows 1 and 5 interchanged prior to the first step.

(Solution on p. 38.)

READ S: page 143, line −5 to page 146, line −5.

Notes

(i) *S: page 144, line 3*
A **band matrix** is one in which all nonzero entries are confined to the leading diagonal and those diagonals close to the leading one. That is, if $a_{i,j}$ is that element of a matrix A which lies in row i and column j, then A is a band matrix with **band width** $2k + 1$ if

$$a_{i,j} = 0 \text{ for } |i - j| > k.$$

For example, a tridiagonal matrix is a band matrix with band width 3 since

$$a_{i,j} = 0 \text{ for } |i - j| > 1.$$

The matrix (1) exhibited earlier in this section has band width 9.

(ii) *S: page 144, line −2 to page 146, line 4*
Gauss elimination with pivoting is the subject of *Unit M201 8*. The idea of triangular decomposition of a matrix was also discussed in that unit.

(iii) *S: page 146, lines 7 and 8*
One of the iterative methods we shall study later in this unit is the *successive over-relaxation method* (SOR for short) in which there is a parameter called the *over-relaxation factor*. The rate of convergence of the SOR method depends on this factor and as a result a large proportion of the theory associated with this method deals with finding a value of the parameter (called the *optimum* over-relaxation factor) which maximizes the rate of convergence of the method.

(iv) *S: page 146, lines 9 to 11*
A system of simultaneous equations can be written in matrix form as

$$A\mathbf{x} = \mathbf{b}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the column vector of unknowns. If we have an approximation $\bar{\mathbf{x}}$ to $\mathbf{x}$ then the vector

$$\mathbf{r} = \mathbf{b} - A\bar{\mathbf{x}}$$

is called the **residual vector** of the given approximation. We can interpret the residual vector as a measure of how close the approximation $\bar{\mathbf{x}}$ comes to satisfying the given system of equations.

For a discussion of residuals and their uses see Section 8.4.1 of *Unit M201 8*.

11.3.2 Theory of Iterative Processes

In this section we shall investigate iterative processes in some generality. This will enable us to find a general result about the convergence of iterative methods which we can then apply to the special methods used for the solution of boundary value problems. The ideas we shall begin with should be familiar to you from work done in both the Mathematics Foundation Course and the Linear Mathematics Course. In particular we adopt here the approach of *Unit M100 28, Linear Algebra IV*; you might like to review Section 28.2 before continuing with the present work.

We begin by considering the system of equations written as

$$A\mathbf{x} = \mathbf{b}. \tag{3}$$

In general we can rearrange this equation as

$$\mathbf{x} = G\mathbf{x} + H\mathbf{b}. \tag{4}$$

For example, adding $\mathbf{x} - A\mathbf{x}$ to each side of Equation (3) yields

$$\mathbf{x} = (I - A)\mathbf{x} + \mathbf{b} \tag{5}$$

where I is the identity matrix. Equation (5) is a special case of Equation (4) where $G = I - A$ and $H = I$. We shall return to Equation (5) later, but for the moment let us look at the iterative scheme which we can obtain from Equation (4) written as

$$\mathbf{x}^{(n)} = G\mathbf{x}^{(n-1)} + H\mathbf{b}. \tag{6}$$

Beginning with a certain initial approximation $\mathbf{x}^{(0)}$, which can be chosen arbitrarily, Equation (6) yields successive approximations $\mathbf{x}^{(n)}$ $(n = 1, 2, \ldots)$. We would like to determine the conditions under which we can be sure that the successive approximations $\mathbf{x}^{(n)}$ converge and that the limit is a solution of the given Equation (3). The latter requirement is dealt with in our next theorem.

THEOREM 3

If the process defined by Equation (6) converges then $\lim \mathbf{x}^{(n)}$ is a solution of Equation (3).

Proof

Suppose that

$$\lim_{n\to\infty} \mathbf{x}^{(n)} = \mathbf{x}^*.$$

Proceeding to the limit in Equation (6) we obtain

$$\mathbf{x}^* = G\mathbf{x}^* + H\mathbf{b}.$$

That is, $\mathbf{x}^*$ satisfies (4), and hence Equation (3), since (4) is simply a rearrangement of (3).

It is clear that if we choose $\mathbf{x}^*$ as the initial vector $\mathbf{x}^{(0)}$ then all subsequent approximations $\mathbf{x}^{(n)}$ $(n = 1, 2, \ldots)$ will be equal to $\mathbf{x}^*$. This fact leads us to call $\mathbf{x}^*$ a *fixed point of the iteration*. Whilst we are on the subject of terminology, we call iterative schemes of the type (6) **stationary** processes whenever the matrices G and H remain constant throughout the whole computation. (Non-stationary processes have been constructed with the general form

$$\mathbf{x}^{(n)} = G^{(n)}\mathbf{x}^{(n-1)} + H^{(n)}\mathbf{b}$$

where the matrices $G^{(n)}$ and $H^{(n)}$ depend on n. In this course we shall deal only with stationary processes of the type (6).)

Turning now to the conditions under which Equation (6) converges we state the following theorem.

THEOREM 4

For a stationary iterative process given by

$$\mathbf{x}^{(n)} = G\mathbf{x}^{(n-1)} + H\mathbf{b}$$

to converge for any initial vector $\mathbf{x}^{(0)}$ it is necessary and sufficient that all eigenvalues of the matrix G be less than one in modulus. That is, the spectral radius of G must be less than unity. (See *Unit 8* for the definition of spectral radius.)

We shall not prove this theorem for an arbitrary $m \times m$ matrix G. Instead, we shall verify it for the case when G has m linearly independent eigenvectors. The theorem is also true for a general square matrix, but the proof takes too long even though it requires no knowledge of linear algebra beyond that contained in M201. In addition we shall simplify matters by considering a process derived from

$$A\mathbf{x} = \mathbf{b}$$

only for the case when A is nonsingular, so that the solution is unique.

Proof

Consider the process

$$\mathbf{x}^{(n)} = G\mathbf{x}^{(n-1)} + H\mathbf{b},$$

and let $\mathbf{x}^*$ be the solution of Equation (3), so that

$$\mathbf{x}^* = G\mathbf{x}^* + H\mathbf{b}.$$

Subtracting the previous equation we obtain

$$\mathbf{x}^* - \mathbf{x}^{(n)} = G(\mathbf{x}^* - \mathbf{x}^{(n-1)}).$$

The vector $\mathbf{e}^{(n)} = \mathbf{x}^* - \mathbf{x}^{(n)}$ represents the error* in the nth approximation to $\mathbf{x}^*$ and the last equation gives

$$\mathbf{e}^{(n)} = G\mathbf{e}^{(n-1)} = G^2\mathbf{e}^{(n-2)} = \cdots = G^n\mathbf{e}^{(0)}.$$

If we suppose that $\{\mathbf{x}^{(n)}\}$ converges as $n \rightsquigarrow \infty$, its limit is $\mathbf{x}^*$ by Theorem 3 and it follows that $\lim \mathbf{e}^{(n)} = \mathbf{0}$ as $n \rightsquigarrow \infty$; hence

$$\lim G^n\mathbf{e}^{(0)} = \mathbf{0}.$$

Since this must hold for any initial vector $\mathbf{x}^{(0)}$, and so for any error vector $\mathbf{e}^{(0)}$, it is necessary that G^n approaches 0, the zero matrix. Conversely if

$$\lim_{n \rightsquigarrow \infty} G^n$$

is the zero matrix it is clear that $\{\mathbf{x}^{(n)}\}$ converges to $\mathbf{x}^*$.

We now assume that the matrix G has m linearly independent eigenvectors which therefore may be used as a basis in the m-dimensional space of column vectors. Therefore, we may write

$$\mathbf{e}^{(0)} = \sum_{i=1}^{m} c_i\mathbf{v}_i$$

where the $\mathbf{v}_i$ are the eigenvectors of G and the c_i are scalars. Since

$$\mathbf{e}^{(n)} = G^n\mathbf{e}^{(0)}$$

we see that

$$\mathbf{e}^{(n)} = \sum_{i=1}^{m} c_i G^n\mathbf{v}_i.$$

Now, by the definition of an eigenvalue we have

$$G^n\mathbf{v}_i = \lambda_i^n\mathbf{v}_i$$

* In *Unit M100 28* the error vector was defined as the negative of this one.

where λ_i is the eigenvalue corresponding to $\mathbf{v}_i$. Therefore, for $\lim \mathbf{e}^{(n)} = \mathbf{0}$ as $n \rightsquigarrow \infty$ we require

$$\mathbf{e}^{(n)} = \sum_{i=1}^{m} c_i \lambda_i^n \mathbf{v}_i \text{ approaches } \mathbf{0} \text{ as } n \rightsquigarrow \infty,$$

which can happen for arbitrary $\mathbf{e}^{(0)}$ if and only if $|\lambda_i| < 1$ for all $i = 1, 2, \ldots, m$. Hence, $\lim G^n = 0$ if and only if $|\lambda_i| < 1$ for all i, that is, if the spectral radius of G is less than unity.

Tying up the two parts of the proof, we have proved our theorem for this special case.

SAQ 6

Use the simple iterative scheme given by Equation (5) to solve the system of equations

$$\begin{aligned} 0.5x_1 - 0.8x_2 &= \quad 1.7, \\ -0.2x_1 + 0.5x_2 &= -0.5, \end{aligned}$$

with the initial approximation $(x_1, x_2) = (0, 0)$, performing the first five iterations only. What can you say from your results about the effectiveness of the simple iterative scheme in solving the given system of equations? Verify that the vector (5, 1) is a fixed point of the iteration.

(Solution on p. 39.)

SAQ 7

Can the system of equations

$$\begin{aligned} x_1 + \quad x_2 &= 4 \\ 3x_1 + 7x_2 &= 20 \end{aligned}$$

be solved by the iterative method given by Equation (5)?

(Solution on p. 40.)

11.3.3 Convergence Rates

We have seen in SAQ 6 that the simple iterative scheme

$$\mathbf{x}^{(n)} = (I - A)\mathbf{x}^{(n-1)} + \mathbf{b}$$

can converge very slowly. Since we can construct many different iterative schemes of the type

$$\mathbf{x}^{(n)} = G\mathbf{x}^{(n-1)} + H\mathbf{b}$$

by choosing different matrices G and H it seems reasonable to choose the one which is most rapidly convergent for a particular problem. In order to be able to compare schemes we need a measure of the rate of convergence of any given choice.

Once again if we assume that the matrix G has m linearly independent eigenvectors $\mathbf{v}_i$ corresponding to eigenvalues λ_i we obtain, for some choice of constants c_i, the relationship

$$\mathbf{e}^{(n)} = \sum_{i=1}^{m} c_i \lambda_i^n \mathbf{v}_i \tag{7}$$

where $\mathbf{e}^{(n)}$ is the difference between the true solution of the system and the nth approximation as defined in Section 11.3.2. For simplicity we shall assume that λ_1 is the unique eigenvalue of largest modulus, i.e. $|\lambda_1| > |\lambda_i|$ for all $i \neq 1$; thus the ratios $(\lambda_i/\lambda_1)^n$ approach zero as n becomes large. Therefore, writing (7) in the form

$$\mathbf{e}^{(n)} = \lambda_1^n \left\{ c_1\mathbf{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^n c_2\mathbf{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^n c_3\mathbf{v}_3 + \cdots + \left(\frac{\lambda_m}{\lambda_1}\right)^n c_m\mathbf{v}_m \right\}$$

we see that

$$\mathbf{e}^{(n)} \simeq \lambda_1^n c_1 \mathbf{v}_1 \qquad \text{for } n \text{ large.}$$

Similarly

$$\mathbf{e}^{(n+1)} \simeq \lambda_1^{n+1} c_1 \mathbf{v}_1 \qquad \text{for } n \text{ large.}$$

and therefore $\mathbf{e}^{(n+1)} \simeq \lambda_1 \mathbf{e}^{(n)}$ for sufficiently large n. As $|\lambda_1|$, by definition, is the *spectral radius* ρ of G, we can say that the error in the approximation ultimately decreases approximately by a factor $1/\rho$ in each subsequent iteration.

Suppose that the error in the kth component of $\mathbf{x}^{(n)}$ is ε, which is given by the approximation

$$\varepsilon = |e_k^{(n)}| \simeq |\lambda_1^n c_1 v_{k,1}| = \rho^n |\phi| \text{ say,}$$

where $v_{k,1}$ is the kth component of the eigenvector $\mathbf{v}_1$ associated with the eigenvalue λ_1 and $\phi = c_1 v_{k,1}$. Taking logarithms we have

$$n = \frac{\log(|\phi|/\varepsilon)}{-\log \rho}.$$

Hence n, the number of iterations required to reduce to ε the error in each component of the solution vector, is inversely proportional to $-\log \rho$. We therefore define the **asymptotic rate of convergence** of the iterative method as $-\log \rho$, and in the choice of a good iterative method our aim is to select a matrix G whose spectral radius is as small as possible so that the rate of convergence is as large as possible.

SAQ 8

What is the asymptotic rate of convergence of the scheme used in SAQ 6?

(Solution on p. 40.)

We have assumed in our analysis that $|\lambda_1| > |\lambda_i|$ for $i \neq 1$. Since G is a real matrix its complex eigenvalues occur in conjugate pairs, which must have the same modulus; hence λ_1 is real. In this case we can approach the question of the rate of convergence somewhat differently and in the process compute an approximation to λ_1. Suppose that $\Delta^{(n)}$ is the correction, or **displacement vector**, to $\mathbf{x}^{(n)}$ at the $(n+1)$th iteration. That is, we define

$$\Delta^{(n)} = \mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}.$$

The iterative scheme can then be regarded as the summation of the infinite series

$$\mathbf{x}^* = \mathbf{x}^{(n)} + \Delta^{(n)} + \Delta^{(n+1)} + \cdots.$$

Now we have seen that for sufficiently large n the error vector $\mathbf{e}^{(n)} = \mathbf{x}^* - \mathbf{x}^{(n)}$ satisfies the relation

$$\mathbf{e}^{(n+1)} \simeq \lambda_1 \mathbf{e}^{(n)}$$

and it is easy to see, by similar reasoning, that the displacement vector satisfies a similar equation given by

$$\Delta^{(n+1)} \simeq \lambda_1 \Delta^{(n)}.$$

(This can hold only if λ_1 is real.) Thus, for sufficiently large n

$$\mathbf{x}^* \simeq \mathbf{x}^{(n)} + \Delta^{(n)}(1 + \lambda_1 + \lambda_1^2 + \lambda_1^3 + \cdots),$$

and if $|\lambda_1| < 1$ we can sum the infinite geometric series and obtain

$$\mathbf{x}^* \simeq \mathbf{x}^{(n)} + \frac{\Delta^{(n)}}{1 - \lambda_1}. \tag{8}$$

This result tells us that the current correction $\Delta^{(n)}$ should really be multiplied by $(1 - \lambda_1)^{-1}$.

This is an important result because if we want an approximation to $\mathbf{x}^*$ with error no greater than ε we might be tempted to terminate the iteration at the first n for which $\|\Delta^{(n)}\| \leqslant \varepsilon$, where the $\| \ \|$ notation denotes the uniform norm,

$$\|\Delta\| = \max_{1 \leq k \leq m} |\Delta_k|.$$

If λ_1 is 0.99, for example, which is not uncommon, such a termination would give a very poor result, and the iteration should be continued until $\|\Delta^{(n)}\| \simeq 0.01\,\varepsilon$. In general we do not know λ_1 but it can be estimated from the ultimate ratio of the corresponding components of successive displacement vectors which we can easily compute. Such information is very important because it not only tells us when to stop the iteration but it can also help us to accelerate the convergence. This is demonstrated in the following SAQ.

SAQ 9

From the results of your iteration in SAQ 6, estimate the eigenvalue of largest modulus of the iteration matrix. Using this estimate, try to compute a better approximation to $\mathbf{x}^*$ from the approximations $\mathbf{x}^{(4)}$ and $\mathbf{x}^{(5)}$.

HINT: Consider Equation (8).

(Solution on p. 40.)

11.4 ITERATIVE METHODS FOR BOUNDARY VALUE PROBLEMS

11.4.1 Jacobi and Gauss-Seidel Methods

This section is intended as an introduction to one class of iterative schemes arising naturally in the solution of the finite-difference equations which approximate elliptic partial differential equations. The methods we are describing here can also be used for the solution of *initial value* problems. For these, however, direct methods of solution (Gauss elimination) are far more useful, and it is for boundary value problems that iterative methods become really important.

Throughout Section 11.4 we shall investigate the solution of Poisson's equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + f = 0$$

in a rectangular region. We have rewritten Poisson's equation in this form to conform with the usage in the next reading passage from *S*. Furthermore we shall discretize the problem using a square mesh with equal mesh spacings h in both the x- and y-directions.* We shall assume that Dirichlet boundary conditions are specified, with U given at all points on the boundary.

In the following reading passage many references are made to passages in Chapter 3 of *S* which you have not read. You need not refer to Chapter 3 at all since the notes in this text contain all the extra information you will require.

*READ S: page 146, line −4, **Systematic iterative methods** to page 149, line 7.*

Notes

(i) *S: page 147, **Jacobi method***
If we write the Jacobi method in matrix form we obtain

$$\mathbf{u}^{(n+1)} = J\mathbf{u}^{(n)} + \mathbf{b}$$

where J is the square matrix of order $(p-1)(q-1)$ given by

$$J = \tfrac{1}{4}\begin{bmatrix} B & I & & & & & \\ I & B & I & & & & \\ & I & B & I & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & I & B & I \\ & & & & & I & B \end{bmatrix}$$

in which each block is a square matrix of order $(q-1)$, with

$$B = \begin{bmatrix} 0 & 1 & & & & & \\ 1 & 0 & 1 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 1 & 0 & 1 & \\ & & & & 1 & 0 & \end{bmatrix}$$

* This presumes that the ratio of the lengths of adjacent sides of the domain is rational.

and I the identity matrix. The column vector $\mathbf{b}$ is determined by the $f_{i,j}$ and $b_{i,j}$. We call J the **Jacobi iteration matrix**. We can see immediately that the iterative scheme is in the form of Equation (6) of Section 11.3.2 and therefore to investigate its convergence we need to find the spectral radius of the matrix J.

In fact, as shown in the Appendix, the eigenvalues λ of J are given by

$$\lambda_{ij} = \frac{1}{4}\left(2\cos\frac{i\pi}{p} + 2\cos\frac{j\pi}{q}\right) \qquad i = 1, 2, \ldots, p-1, j = 1, 2, \ldots, q-1,$$

and the spectral radius ρ is therefore

$$\rho(J) = \frac{1}{2}\left(\cos\frac{\pi}{p} + \cos\frac{\pi}{q}\right).$$

Hence $\rho(J) < 1$ (since $p, q > 1$) and the Jacobi method is convergent.

(ii) *S: pages 147 to 149, **Gauss-Seidel method***
In matrix form Equation (5.7) of *S: page 148* becomes

$$\mathbf{u}^{(n+1)} = L\mathbf{u}^{(n+1)} + U\mathbf{u}^{(n)} + \mathbf{b},$$

where

$$L = \tfrac{1}{4}\begin{bmatrix} B_L & & & & & \\ I & B_L & & & & \\ & I & B_L & & & \\ & & \ddots & \ddots & & \\ & & & I & B_L & \\ & & & & I & B_L \end{bmatrix}, \quad B_L = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \\ & & & & 1 & 0 \end{bmatrix},$$

and

$$U = \tfrac{1}{4}\begin{bmatrix} B_U & I & & & & \\ & B_U & I & & & \\ & & B_U & I & & \\ & & & \ddots & \ddots & \\ & & & & B_U & I \\ & & & & & B_U \end{bmatrix}, \quad B_U = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix}.$$

L and U are strictly lower and strictly upper triangular matrices respectively of order $(p-1)(q-1)$, B_L and B_U are square matrices of order $q-1$ and I is the unit matrix of order $q-1$. Notice that $L + U = J$, the Jacobi iteration matrix. The Gauss-Seidel method for our problem can be rewritten as

$$(I - L)\mathbf{u}^{(n+1)} = U\mathbf{u}^{(n)} + \mathbf{b}$$

and we see that

$$\mathbf{u}^{(n+1)} = (I - L)^{-1}U\mathbf{u}^{(n)} + (I - L)^{-1}\mathbf{b}$$

which is of the general form of the iterative schemes dealt with in Section 11.3. We call the matrix $(I - L)^{-1}U$ the **Gauss-Seidel iteration matrix**. Fortunately, for the matrices which we are considering, it is not necessary to find the eigenvalues of the Gauss-Seidel iteration matrix directly since it is possible to express them in terms of the eigenvalues of the Jacobi iteration matrix. We shall not do this here. However, notice the relationship, given at the top of *S: page 149*, between the spectral radii of the two iteration matrices.

General Comment

We may use an iterative scheme to solve the matrix equation

$$A\mathbf{x} = \mathbf{b}$$

where

$$A = I - L - U,$$

with I the identity matrix and L, U being strictly lower and strictly upper triangular matrices respectively. In general we call the scheme

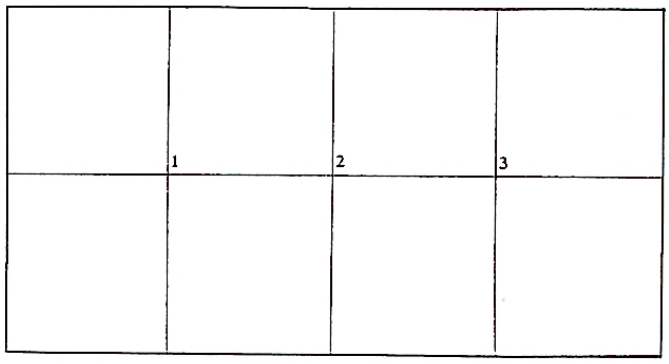
$$\mathbf{x}^{(n+1)} = L\mathbf{x}^{(n)} + U\mathbf{x}^{(n)} + \mathbf{b}$$

a *Jacobi method*, and

$$\mathbf{x}^{(n+1)} = L\mathbf{x}^{(n+1)} + U\mathbf{x}^{(n)} + \mathbf{b}$$

a *Gauss-Seidel method*.

SAQ 10



The five-point formula, when applied to Laplace's equation for the region shown with Dirichlet boundary conditions, yields the system

$$\begin{aligned} u_1 - \tfrac{1}{4}u_2 &= b_1, \\ -\tfrac{1}{4}u_1 + u_2 - \tfrac{1}{4}u_3 &= b_2, \\ -\tfrac{1}{4}u_2 + u_3 &= b_3. \end{aligned}$$

(a) Determine the Jacobi iteration matrix J.

(b) Determine the Gauss-Seidel iteration matrix G.

(c) Verify that the spectral radii are related by

$$[\rho(J)]^2 = \rho(G),$$

for this problem.

(Solution on p. 40.)

SAQ 11

(a) Show that the Jacobi method is convergent and the Gauss-Seidel method divergent when applied to the system

$$\begin{aligned} u_1 + 2u_2 - 2u_3 &= 1, \\ u_1 + u_2 + u_3 &= 3, \\ 2u_1 + 2u_2 + u_3 &= 5. \end{aligned}$$

(b) What do you find when you perform the Jacobi iteration with a starting value given by

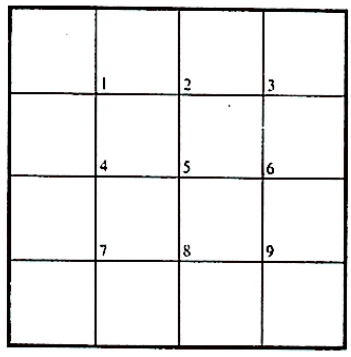
$$\mathbf{x}^{(0)} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$$

and exact arithmetic is used? How could you have predicted your result?

(Solution on p. 42.)

11.4.2 Accelerating the Rates of Convergence

Consider the finite-difference solution of Poisson's equation in the region shown in the figure when Dirichlet conditions are given on the boundary.



If the mesh lengths are equal, the five-point formula produces a set of linear equations of the general form

$$u_{i,j} = \tfrac{1}{4}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) + b_{i,j}.$$

and typical equations relevant to the figure are

$$\begin{aligned} u_1 &= \tfrac{1}{4}(u_2 + u_4) + b_1, \\ u_2 &= \tfrac{1}{4}(u_1 + u_3 + u_5) + b_2, \\ u_5 &= \tfrac{1}{4}(u_2 + u_4 + u_6 + u_8) + b_5, \\ &\text{etc.} \end{aligned} \tag{1}$$

Now the essence of an iterative method is that we scan the mesh points systematically. Having reached point r we obtain a new value for u_r using the relevant equation from the set (1). With this new value of u_r, some of the other equations fail to be satisfied, whence the need for iteration.

The Jacobi and Gauss-Seidel methods perform the operations slightly differently from each other. Suppose that, for example, we start with some initial guessed values $u_1^{(0)}, u_2^{(0)}, \ldots, u_9^{(0)}$ relevant to the figure. At point 1 both methods produce the new approximation $u_1^{(1)}$ from the formula

$$u_1^{(1)} = \tfrac{1}{4}(u_2^{(0)} + u_4^{(0)}) + b_1.$$

However, at the next stage, the adjustment at point 2, the Jacobi method uses the formula

$$u_2^{(1)} = \tfrac{1}{4}(u_1^{(0)} + u_3^{(0)} + u_5^{(0)}) + b_2, \tag{2}$$

while the Gauss-Seidel formula is

$$u_2^{(1)} = \tfrac{1}{4}(u_1^{(1)} + u_3^{(0)} + u_5^{(0)}) + b_2.$$

Thus, for subsequent points, we do not use the new (and hopefully better) value $u_1^{(1)}$ in the Jacobi method, whereas in the Gauss-Seidel case we take advantage of this better available value in order, again hopefully, to accelerate the convergence of the process. If we perform the adjustments scanning the points in the natural (numerical) ordering shown in the figure we see that, for example, in the Gauss-Seidel method the adjustment at point 5 is given by

$$u_5^{(1)} = \tfrac{1}{4}(u_2^{(1)} + u_4^{(1)} + u_6^{(0)} + u_8^{(0)}) + b_5, \tag{3}$$

which uses the two "better" values already known for points 2 and 4.

The rate of convergence of the Jacobi iteration is clearly independent of the order in which the mesh points are scanned since the arithmetic is independent of the ordering. For the Gauss-Seidel method, however, not all orderings will involve the same arithmetic. For example, suppose we scan the points in the figure in the order

1, 2, 3; 6, 5, 4; 7, 8, 9;

then the adjustment at point 5 is given by

$$u_5^{(1)} = \tfrac{1}{4}(u_2^{(1)} + u_4^{(0)} + u_6^{(1)} + u_8^{(0)}) + b_5 \tag{4}$$

since we already have the new values at points 1, 2, 3 and 6 but not at 4, 7, 8 and 9. Equation (4) is obviously different from Equation (3) and we might expect that one of the orderings will have a faster convergence rate than the other.

In discussing this topic it is convenient to consider at the same time the idea of introducing some parameter into the matrix equations, hoping to obtain even faster convergence with the choice of a suitable numerical value for this parameter.

The motivation for this was the discovery by early workers in this field that quite commonly all the successive changes in any pivotal value in successive steps of, say, the Gauss-Seidel process, are of the same sign. For example, approximations at a particular point may have successive values like 100, 120, 130, 135, 137, 138, converging monotonically to the solution. (SAQ 6 is a typical example.) This indeed *must* happen eventually if the eigenvalue λ_1 of largest modulus in the iteration matrix is real and positive since, as we saw in Section 11.3.3, successive changes are ultimately connected by the equation

$$\Delta^{(n+1)} \simeq \lambda_1 \Delta^{(n)}$$

for sufficiently large n; in practice, a positive λ_1 turns out to be the rule rather than the exception. In such cases one might accelerate the convergence by making a larger change $\Delta^{(n)}$ than is needed to satisfy (temporarily) the relevant equation of the set (1). Now, if the original equations are given by

$$A\mathbf{u} = \mathbf{b}$$

where $A = I - L - U$, L being strictly lower triangular and U strictly upper triangular, then the Gauss-Seidel method is given by

$$\mathbf{u}^{(n+1)} = L\mathbf{u}^{(n+1)} + U\mathbf{u}^{(n)} + \mathbf{b}$$

and the current correction is just

$$\Delta^{(n)} \simeq \mathbf{u}^{(n+1)} - \mathbf{u}^{(n)} = \mathbf{b} + L\mathbf{u}^{(n+1)} + (U - I)\mathbf{u}^{(n)}.$$

If we multiply this correction by ω, where usually $1 \leqslant \omega < 2$, the iterative scheme becomes

$$\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)} = \omega\{\mathbf{b} + L\mathbf{u}^{(n+1)} + (U - I)\mathbf{u}^{(n)}\}$$

or

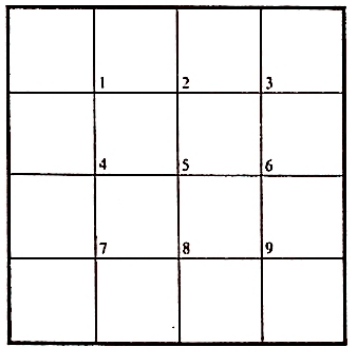
$$(I - \omega L)\mathbf{u}^{(n+1)} = \{(1 - \omega)I + \omega U\}\mathbf{u}^{(n)} + \omega\mathbf{b}. \tag{5}$$

This is the iteration equation for the **successive over-relaxation** or **SOR method** ("over" because at each stage we are "overdoing" things). Notice that the SOR method reduces to the Gauss-Seidel method when $\omega = 1$.

The two questions, of choosing a suitable ordering of the equations and a suitable value for the parameter ω, have received much attention in the last few years, and in the next section we give an introduction, without detailed proofs, to the more important results of this research.

11.4.3 Consistent Ordering, Property *A* and SOR

We consider further Poisson's equation in the square domain of Section 11.4.2.



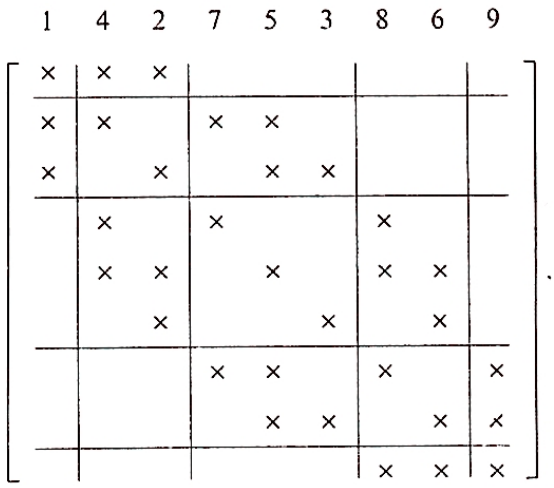
Suppose we order the points for the iterations of Gauss-Seidel or SOR methods along the diagonals, that is, in the order

1; 4, 2; 7, 5, 3; 8, 6; 9.

In matrix form the equations become

$$\mathbf{u}^{(n+1)} = L\mathbf{u}^{(n+1)} + U\mathbf{u}^{(n)} + \mathbf{b}, \tag{6}$$

where $I - L - U$ has the form



in which the crosses represent nonzero entries. Equation (6) implies that we scan the mesh in the diagonal order quoted, adjusting successive values from equations like (1) of the previous section, but using the latest available pivotal values on the right-hand side of (6). For example, Equation (3) for the Gauss-Seidel method becomes

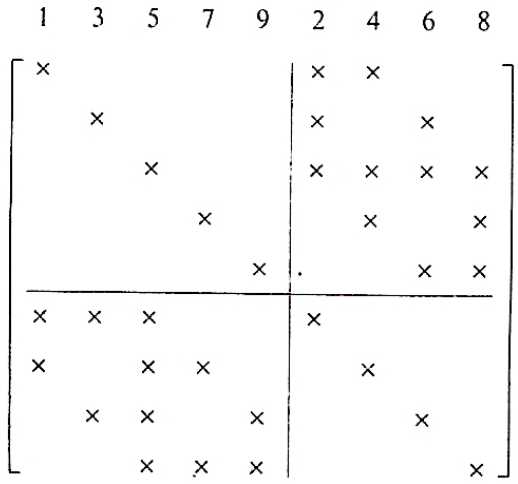
$$u_5^{(n+1)} = \tfrac{1}{4}(u_2^{(n+1)} + u_4^{(n+1)} + u_6^{(n)} + u_8^{(n)}) + b_5, \tag{7}$$

which is of the same form as the natural ordering 1, 2, 3, ..., 9 because in each case u_2 and u_4, but not u_6 and u_8, have already been "adjusted".

Consider next the ordering

1, 3, 5, 7, 9; 2, 4, 6, 8,

the so-called *black-white* or *checker board* ordering. The matrix $I - L - U$ now has the form



and in this case Equation (7) becomes

$$u_5^{(n+1)} = \tfrac{1}{4}(u_2^{(n)} + u_4^{(n)} + u_6^{(n)} + u_8^{(n)}) + b_5$$

since none of the pivotal values on the right-hand side has yet been adjusted. When we come to compute a new $u_2^{(n+1)}$ on the other hand, all the other pivotal values in the relevant equation have the superscript $(n + 1)$.

The partitioned matrix of Equation (6) with the diagonal ordering has the *diagonally block-tridiagonal* form

$$\begin{bmatrix} D_1 & F_1 & & & & & \\ E_1 & D_2 & F_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & E_{m-2} & D_{m-1} & F_{m-1} \\ & & & & E_{m-1} & D_m \end{bmatrix}, \tag{8}$$

where the D_i $(i = 1, 2, \ldots)$ are diagonal matrices, not necessarily of the same order. The partitioned matrix obtained from the checker board ordering, which we can write as

$$\begin{bmatrix} D_1 & F \\ E & D_2 \end{bmatrix}, \tag{9}$$

is of special diagonally block-tridiagonal form.

SAQ 12

What is the form of the matrix obtained using the natural ordering 1, 2, 3; 4, 5, 6; 7, 8, 9? Is it of either of the special forms (8) or (9)?

(Solution on p. 43.)

It turns out that there are other orderings, called **consistent** orderings, which do not produce matrices of types (8) or (9), but for which the arithmetic of the Gauss-Seidel or SOR method is identical with one of these forms. That is, at each stage of the iteration the values obtained at each point are identical with those obtained using one of these forms.

For the diagonal ordering the following table shows, for the computation of successive $u_r^{(n+1)}$, those values in the respective equations which already have the superscript $(n + 1)$ and those which, not yet adjusted in the current iteration cycle, have the superscript (n).

	points with superscript $(n+1)$	*points with superscript* (n)
$r = 1$	none	4, 2
4	1	7, 5
2	1	5, 3
7	4	8
5	4, 2	8, 6
3	2	6
8	7, 5	9
6	5, 3	9
9	8, 6	none

Here the numbers in the second column refer to the elements *below* the diagonal of the matrix corresponding to the row number given in the first column, and the third column refers to the corresponding elements *above* the diagonal of the matrix. The next SAQ shows, among other things, that there is another ordering *not* in the form of (8) but consistent with that of the table above.

SAQ 13

(a) Why are the statements about orderings for the SOR method equivalent to those made about the Gauss-Seidel method?

(b) Show that the natural ordering of SAQ 12 is consistent with that of Equation (6).

(c) Show that the ordering

1, 2, 3; 6, 5, 4; 7, 8, 9

is not consistent with either the diagonal ordering or checker board ordering.

(Solution on p. 43.)

We now give two definitions.

A square matrix is said to have **Property** A if, by transposing pairs of rows and corresponding columns, it can be transformed to the form of the matrix (9). In particular, the matrix (8) has Property A. (To see this, rearrange the matrix so that the order of diagonal blocks is $D_1, D_3, D_5, \ldots, D_2, D_4, \ldots$)

Let Q be a matrix obtained from P by successively interchanging pairs of rows and corresponding columns. For example, let Q be obtained from P by interchanging the ith and jth rows, and interchanging the ith and jth columns. Then, if E_{ij} is obtained from the identity matrix by interchanging the ith and jth rows,

$$Q = E_{ij}^{-1} P E_{ij}.$$

In general, if T is the matrix obtained from the $n \times n$ identity matrix by reordering the rows to

$$i_1, i_2, \ldots, i_n,$$

then

$$Q = T^{-1} P T$$

is the matrix obtained from P by reordering its rows and columns to the new order. It is clear that the main diagonal of Q consists of the elements of the main diagonal of P in the new order. We say that Q is **consistent with** P if the set of nonzero elements below (above) the main diagonal of Q is a permutation of the nonzero elements below (above) the main diagonal of P.

In the context of the Gauss-Seidel method the matrix obtained after reordering $I - L - U$ is consistent with $I - L - U$ if the orderings are consistent in the sense discussed previously. If $I - L - U$ has Property A then, for brevity, we shall say that $I - L - U$ is **consistent** if it is consistent with any matrix of the form (8) *or* (9).

The point of making our definitions is that if $I - L - U$ has Property A, the *asymptotic* rate of convergence of the Gauss-Seidel or SOR methods is the same for *all* orderings such that the matrix $I - L - U$ is *consistent*. It is, of course, not surprising that, for example, the diagonal ordering and the natural ordering should yield the same rate of convergence, since we showed in SAQ 13 that these orderings are *consistent* with each other and give rise to precisely the same arithmetic. The more remarkable thing is that the diagonal ordering and the checker board ordering, which are not consistent with each other and give rise to different arithmetic, should still have the same asymptotic rates of convergence. It is the Property A quality of the original matrix which produces this result.

It might be asked why we concentrate on consistent orderings and on matrices with Property A. First, it can be shown that the asymptotic rate of convergence for all *inconsistent* orderings is slower than that for consistent orderings. Second, for matrices with Property A, and for all consistent orderings, we can find an optimum value for the parameter which gives the fastest rate of convergence for the SOR variation of the Gauss-Seidel method.

For this purpose we need to investigate the spectral radius of the SOR iteration matrix

$$S_\omega = (I - \omega L)^{-1}\{(1 - \omega)I + \omega U\}$$

obtained from Equation (5).

First we shall prove a lemma. (Omit the proof if you are short of time.)

LEMMA

If the matrix $I - L - U$, where I is the identity matrix and $L(U)$ is strictly lower (upper) triangular, has Property A and is consistent then the eigenvalues of $\mu L + \mu^{-1}U$ $(\mu \neq 0)$ are independent of μ.

Proof

Since $I - L - U$ has Property A and is consistent, there exists a matrix T such that

$$T^{-1}(I - L - U)T = \begin{bmatrix} I_1 & F_1 & & & \\ E_1 & I_2 & F_2 & & \\ & \ddots & \ddots & \ddots & \\ & & E_{m-2} & I_{m-1} & F_{m-1} \\ & & & E_{m-1} & I_m \end{bmatrix}.$$

where the elements of $E_1, E_2, \ldots, E_{m-1}$ are the elements of L, and the elements of $F_1, F_2, \ldots, F_{m-1}$ are the elements of U, suitably reordered. (The diagonal blocks are all identity matrices. Why?) Now, it is clear that multiplying I, L or U by a constant will reproduce this multiplication in precisely those elements of the block matrix which come from I, L or U respectively. Hence

$$T^{-1}(\lambda I - \mu L - \mu^{-1}U)T = \begin{bmatrix} \lambda I_1 & \mu^{-1}F_1 & & & \\ \mu E_1 & \lambda I_2 & \mu^{-1}F_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu E_{m-2} & \lambda I_{m-1} & \mu^{-1}F_{m-1} \\ & & & \mu E_{m-1} & \lambda I_m \end{bmatrix}.$$

Now the eigenvalues of $\mu L + \mu^{-1}U$ are the roots of the equation

$$|\lambda I - \mu L - \mu^{-1}U| = 0.$$

Since a similarity transformation preserves the eigenvalues we need look only at the matrix on the right-hand side. Suppose I_i is of order p_i. We leave the first p_1 rows and columns as they stand. Then we multiply the next p_2 rows by μ^{-1} and the next p_2 columns by μ, obtaining the matrix

$$\begin{bmatrix} \lambda I_1 & F_1 & & & \\ E_1 & \lambda I_2 & \mu^{-2}F_2 & & \\ & \mu^2 E_2 & \lambda I_3 & \mu^{-1}F_3 & \\ & & \cdot & \cdot & \cdot \\ & & & \mu E_{m-1} & \lambda I_m \end{bmatrix}.$$

Proceeding in this way, we multiply the next p_3 rows by μ^{-2} and p_3 columns by μ^2, etc., finally multiplying the last p_m rows by μ^{-m+1} and the last p_m columns by μ^{m-1}, leaving us with the result that

$$|\lambda I - \mu L - \mu^{-1}U| = \begin{vmatrix} \lambda I_1 & F_1 & & \\ E_1 & \lambda I_2 & F_2 & \\ & \cdot & \cdot & \cdot \\ & E_{m-2} & \lambda I_{m-1} & F_{m-1} \\ & & E_{m-1} & \lambda I_m \end{vmatrix}$$

$$= |T^{-1}(\lambda I - L - U)T|$$

$$= |\lambda I - L - U|,$$

since the row and column operations have, in total, multiplied the determinant by 1. Hence the eigenvalues of $\mu L + \mu^{-1}U$ are the same as those of $L + U$; i.e., they are independent of μ.

We now use this result to relate the eigenvalues of the SOR iteration matrix to that of the Jacobi iteration matrix.

If λ is an eigenvalue of S_ω then

$$|S_\omega - \lambda I| = 0.$$

Then

$$|(I - \omega L)^{-1}[I + \omega(U - I)] - \lambda I| = 0$$

so that, on multiplying by $|I - \omega L|$, we obtain

$$|I + \omega(U - I) - \lambda(I - \omega L)| = 0$$

since determinants satisfy the rule

$$|PQ| = |P||Q|.$$

Hence

$$|\omega(U + \lambda L) - (\lambda + \omega - 1)I| = 0,$$

i.e.,

$$|\lambda^{\frac{1}{2}}\omega(\lambda^{\frac{1}{2}}L + \lambda^{-\frac{1}{2}}U) - (\lambda + \omega - 1)I| = 0.$$

Finally

$$\left|(\lambda^{\frac{1}{2}}L + \lambda^{-\frac{1}{2}}U) - \frac{\lambda + \omega - 1}{\lambda^{\frac{1}{2}}\omega}I\right| = 0. \tag{10}$$

Now by the Lemma, if $I - L - U$ has Property A and is *consistent*, then the eigenvalues of the matrix $\mu L + \mu^{-1}U$ are independent of μ. It follows that the eigenvalues of $\lambda^{\frac{1}{2}} L + \lambda^{-\frac{1}{2}}U$ are identical with those of $L + U$, and these are precisely the eigenvalues α of the *Jacobi* iteration matrix. Equation (10) then gives the remarkable relation

$$\alpha = \frac{\lambda + \omega - 1}{\lambda^{\frac{1}{2}}\omega} \tag{11}$$

between the eigenvalues α of the Jacobi iteration matrix and the eigenvalues λ of the SOR iteration matrix. With $\omega = 1$, incidentally, we confirm the result $\lambda = \alpha^2$ which we have already observed, for example, in SAQ 10 and which guarantees, when $|\alpha| < 1$ for the convergence of the Jacobi method, that the Gauss-Seidel method converges twice as fast.

Proceeding from Equation (11) we can show that the value of the parameter α which minimizes the spectral radius of the SOR iteration matrix, thereby giving the fastest possible rate of convergence, is given by

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - [\rho(J)]^2}},$$

where $\rho(J)$ is the spectral radius of the Jacobi iteration matrix. Moreover, the spectral radius of the SOR iteration matrix is then given simply by

$$\rho(S_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1.$$

We have developed the program $SOR 321 to solve up to ten equations iteratively by SOR; a listing follows.

```
10  PRINT "THIS PROGRAM  USES THE METHOD OF SUCCESSIVE OVER RELAXATION
20  PRINT "TO SOLVE A SYSTEM OF LINEAR EQUATIONS TO SIX DEC.PL."
30  PRINT
40  DIM A[10,11],X[10],Z[10],A$[3]
50  PRINT "TYPE NUMBER OF EQUATIONS(<=10)";
60  INPUT N
70  PRINT
80  PRINT "TYPE ELEMENTS OF COEFFICIENT MATRIX,ROW BY ROW"
90  FOR I=1 TO N
100  PRINT "ROW";I;
110  MAT  INPUT Z[N]
120  FOR J=1 TO N
130  LET A[I,J]=Z[J]
140  NEXT J
150  NEXT I
160  PRINT
170  PRINT "TYPE R.H.S. CONSTANTS"
180  MAT  INPUT Z[N]
190  FOR I=1 TO N
200  LET A[I,N+1]=Z[I]
210  NEXT I
220  PRINT
230  PRINT "TYPE INITIAL APPROXIMATION"
240  MAT  INPUT Z[N]
250  PRINT
260  PRINT "TYPE OVER RELAXATION FACTOR";
270  INPUT W
280  MAT X=ZER[N]
290  PRINT
300  PRINT
310  LET C=0
320  FOR I=1 TO N
```

```
330  LET M=A[I,I]
340  FOR J=1 TO N+1
350  LET A[I,J]=A[I,J]/M
360  NEXT J
370  NEXT I
380  LET C=C+1
390  LET E=0
400  MAT X=Z
410  FOR I=1 TO N
420  LET S1=S2=0
430  FOR J=I+1 TO N
440  LET S1=S1+A[I,J]*X[J]
450  NEXT J
460  FOR J=1 TO I-1
470  LET S2=S2+A[I,J]*Z[J]
480  NEXT J
490  LET Z[I]=(1-W)*X[I]+W*(A[I,N+1]-S1-S2)
500  LET K=ABS(Z[I]-X[I])
510  IF K <= 5.E-07 THEN 530
520  LET E=1
530  NEXT I
540  IF E=1 THEN 380
550  PRINT
560  PRINT "SOLUTION"
570  PRINT "--------"
580  MAT  PRINT Z
590  PRINT
600  PRINT "NUMBER OF ITERATIONS NEEDED IS";C
610  PRINT
620  PRINT "REPEAT WITH ANOTHER OVER-RELAXATION FACTOR/INITIAL APPROX.
630  PRINT "YES OR NO";
640  INPUT A$
650  IF A$="YES" THEN 230
660  END
```

SAQ 14

Consider the equations of SAQ 10, and its solution which gives

$$[\rho(J)]^2 = \tfrac{1}{8} = \rho(G).$$

(a) Find the optimum ω for the SOR method and the spectral radius of the SOR iteration matrix. Does the original matrix have Property A?

(b) Determine the improvement in convergence by performing both Gauss-Seidel and SOR iterations with the optimum value of ω, for the problem of SAQ 10 with $b_1 = 1.5, b_2 = 3, b_3 = 3.5$. You may use the library program $SOR321.

(Solution on p. 44.)

Two questions remain. Firstly, to get ω_{opt} we apparently need an estimate of $\rho(J)$. This is sometimes known theoretically, as in note (i) of Section 11.4.1. Alternatively we iterate for some time with the Gauss-Seidel method, computing the spectral radius of its iteration matrix, which is $[\rho(J)]^2$, by the method suggested at the end of Section 11.3.3. In practice, it is better to overestimate the spectral radius of the Jacobi iteration matrix because it will yield a smaller decrease in the rate of convergence of the SOR method than would a comparable underestimate.

Secondly, what can we do if our matrix does not have Property A? In such cases the SOR method has been used with some success, though without theoretical justification, and much current research is devoted to this problem.

11.5 SUMMARY

We began the unit by investigating two simple applications of the finite-difference method applied to boundary value problems. We saw that the numerical method requires investigation in two different areas. The finite-difference method gives rise to (large) systems of linear algebraic equations and we need to know the conditions under which the solution of the equations converges to the true solution of the boundary value problem. We were able to prove, for bounded regions, that the *five-point formula* for the Laplacian is convergent and moreover we were able to obtain a bound on the error of our approximate solution. Both of these results are consequences of a *maximum principle* which we also established.

The second half of the unit concentrated on the problem of solving the systems of algebraic equations which arise in the finite-difference method. In particular, we illustrated that iterative methods have advantages over other methods for our particular problems. We investigated iterative methods and showed that in general the iteration matrix must have a spectral radius less than unity, and that the smaller the spectral radius the faster is the *asymptotic rate of convergence* of a method.

In the final section we looked at some particular iterative methods for boundary value problems, namely the *Jacobi*, *Gauss-Seidel* and *SOR* methods. We found that the SOR method could be optimized to give the fastest rate of convergence of these methods. However, our analysis depended upon the initial matrix having *Property A* and the mesh points being ordered in a manner *consistent* with this property. Under these circumstances we were able to state a result giving the *optimum relaxation factor* yielding the fastest rate of convergence.

An alternative treatment of some of the material presented in Sections 11.3 and 11.4 may be found in *S: pages 76 to 79.*

11.6 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

We are given

$$\frac{\partial^2 u}{\partial x^2} + 3\frac{\partial^2 u}{\partial y^2} = -16 \qquad -1 \leqslant x \leqslant 1,\ -1 \leqslant y \leqslant 1,$$

$$u(1, y) = 0 \qquad -1 \leqslant y \leqslant 1,$$

$$\frac{\partial u}{\partial y}(x, 1) = -u(x, 1) \qquad -1 \leqslant x \leqslant 1.$$

Putting $y = 3^{\frac{1}{4}}\bar{y}$, we obtain

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial \bar{y}^2} = -16 \qquad -1 \leqslant x \leqslant 1,\ -\frac{1}{3^{\frac{1}{4}}} \leqslant \bar{y} \leqslant \frac{1}{3^{\frac{1}{4}}},$$

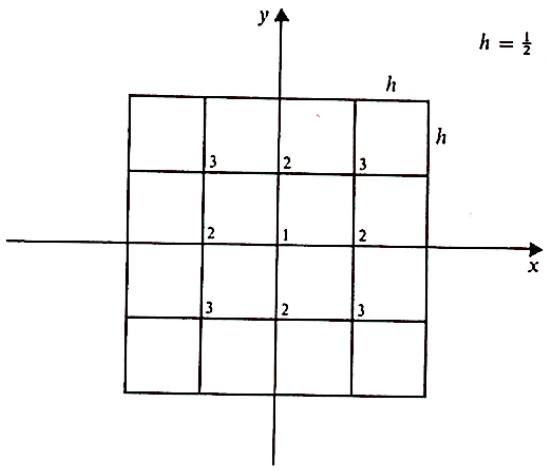
$$u(1, \bar{y}) = 0 \qquad -\frac{1}{3^{\frac{1}{4}}} \leqslant \bar{y} \leqslant \frac{1}{3^{\frac{1}{4}}},$$

$$\frac{\partial u}{\partial \bar{y}}(x, 3^{\frac{1}{4}}) = -3^{\frac{1}{4}}u(x, 3^{\frac{1}{4}}) \qquad -1 \leqslant x \leqslant 1.$$

If we take a mesh spacing of $\frac{1}{4}$ in the x-direction and $\frac{1}{4\sqrt{3}}$ in the $\bar{y}$-direction, the central-difference formula is the same as that given on the last line of *S: page 134*, and the numerical results remain the same.

Solution to SAQ 2

See the solution given on *S: page 162, lines 5 to 8*. The following figure may be of assistance.



Solution to SAQ 3

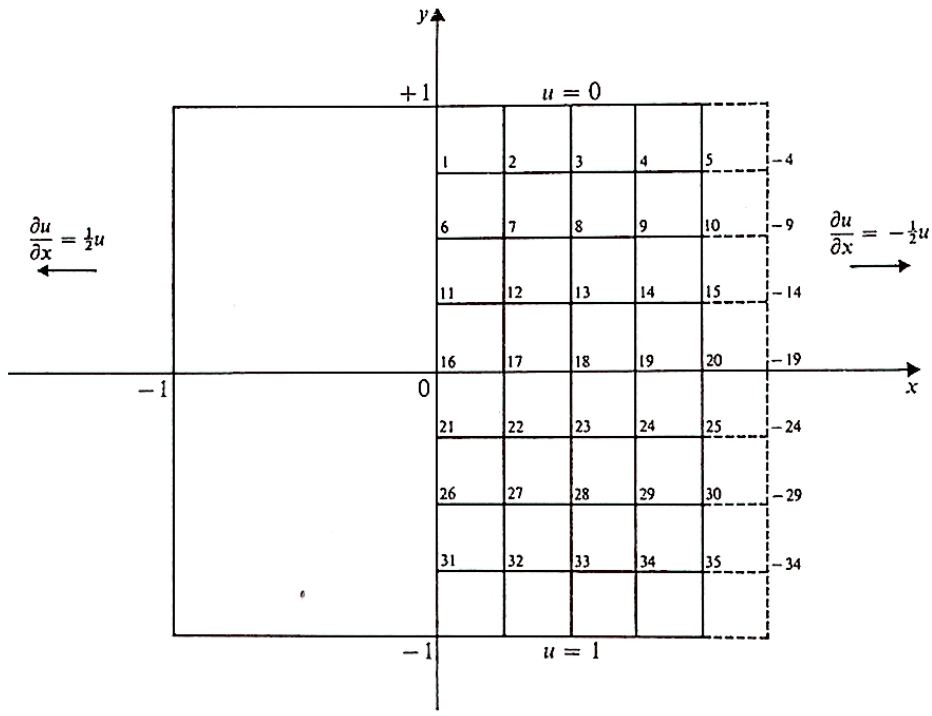
The differential equation is approximated by

$$\frac{1}{h^2}(\delta_x^2 + \delta_y^2)u_{i,j} - 32u_{i,j} = 0$$

and the derivative boundary condition along $x = 1$ by

$$\frac{1}{2h}(u_{i+1,j} - u_{i-1,j}) = -\frac{1}{2}u_{i,j}.$$

The region is shown in the following diagram which also shows the mesh point numbering and the extension of the mesh to enable the derivative boundary condition to be dealt with by the central-difference formula.



We see that there is symmetry about the y-axis and therefore taking $h = \frac{1}{4}$ there are 35 unknowns to be evaluated.

Typical equations are:

for point 1,

$$2u_2 + u_6 - 6u_1 = 0,$$

for point 2,

$$u_1 + u_3 + u_7 - 6u_2 = 0,$$

for point 3,

$$u_2 + u_4 + u_8 - 6u_3 = 0,$$

for point 4,

$$u_3 + u_5 + u_9 - 6u_4 = 0,$$

for point 5 (two equations since it is a boundary point),

$$u_4 + u_{-4} + u_{10} - 6u_5 = 0$$

$$u_{-4} - u_4 = -\tfrac{1}{4}u_5$$

and eliminating u_{-4} gives

$$2u_4 + u_{10} - 6\tfrac{1}{4}u_5 = 0.$$

Repeating this process for the remaining points yields

$$\left[\begin{array}{cccccc|cccccc|cc}
-6 & 2 & & & & & 1 & & & & & & & \\
1 & -6 & 1 & & & & & 1 & & & & & & \\
 & 1 & -6 & 1 & & & & & 1 & & & & & \\
 & & 1 & -6 & 1 & & & & & 1 & & & & \\
 & & & 2 & -6\frac{1}{4} & & & & & & 1 & & & \\
\hline
1 & & & & & & -6 & 2 & & & & 1 & & \\
 & 1 & & & & & 1 & -6 & 1 & & & & 1 & \\
 & & 1 & & & & & 1 & -6 & 1 & & & & 1 \\
 & & & 1 & & & & & 1 & -6 & 1 & & & \\
 & & & & 1 & & & & & 2 & -6\frac{1}{4} & & & 1 \\
\hline
\vdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \vdots &
\end{array}\right]
\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ \vdots \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

which gives the required form. Note that the constant vector **b** has -1's in the last five positions due to the boundary condition $u = 1$ along $y = -1$.

Solution to SAQ 4

The local truncation error $T_{i,j}$ of the five-point formula is given by

$$T_{i,j} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{(\Delta x)^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{(\Delta y)^2} - f_{i,j}.$$

By Taylor's Theorem

$$U_{i+1,j} - 2U_{i,j} + U_{i-1,j} = (\Delta x)^2 \frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{(\Delta x)^4}{12} \frac{\partial^4 U_{i,j}}{\partial x^4} + O((\Delta x)^6),$$

$$U_{i,j+1} - 2U_{i,j} + U_{i,j-1} = (\Delta y)^2 \frac{\partial^2 U_{i,j}}{\partial y^2} + \frac{(\Delta y)^4}{12} \frac{\partial^4 U_{i,j}}{\partial y^4} + O((\Delta y)^6).$$

Hence,

$$T_{i,j} = \frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{\partial^2 U_{i,j}}{\partial y^2} - f_{i,j} + \frac{1}{12}\left[(\Delta x)^2 \frac{\partial^4 U_{i,j}}{\partial x^4} + (\Delta y)^2 \frac{\partial^4 U_{i,j}}{\partial y^4}\right] + O((\Delta x)^4) + O((\Delta y)^4).$$

Since the differential equation gives

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f,$$

we find that

$$T_{i,j} = \frac{1}{12}\left[(\Delta x)^2 \frac{\partial^4 U_{i,j}}{\partial x^4} + (\Delta y)^2 \frac{\partial^4 U_{i,j}}{\partial y^4}\right] + O((\Delta x)^4) + O((\Delta y)^4).$$

Taylor's Theorem with remainder gives

$$U_{i+1,j} = U_{i,j} + h\frac{\partial U_{i,j}}{\partial x} + \frac{h^2}{2}\frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 U_{i,j}}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 U}{\partial x^4}(x_i + \theta_1 h, y_j)$$

for some $\theta_1 \in (0, 1)$,

where $h = \Delta x$. Also,

$$U_{i-1,j} = U_{i,j} - h\frac{\partial U_{i,j}}{\partial x} + \frac{h^2}{2}\frac{\partial^2 U_{i,j}}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 U_{i,j}}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 U}{\partial x^4}(x_i - \theta_2 h, y_j)$$

for some $\theta_2 \in (0, 1)$,

and so,

$$U_{i+1,j} - 2U_{i,j} + U_{i-1,j} = h^2\frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 U}{\partial x^4}(x_i + \theta_3 h, y_j)$$

for some $\theta_3 \in (-1, 1)$,

using the Intermediate Value Theorem given in note (ii) of Section 8.2.1 of *Unit 8*.

Similarly, on putting $k = \Delta y$, we obtain

$$U_{i,j+1} - 2U_{i,j} + U_{i,j-1} = k^2\frac{\partial^2 U_{i,j}}{\partial y^2} + \frac{k^4}{12}\frac{\partial^4 U}{\partial y^4}(x_i, y_j + \theta_4 k)$$

for some $\theta_4 \in (-1, 1)$.

Therefore,

$$T_{i,j} = \frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{\partial^2 U_{i,j}}{\partial y^2} - f_{i,j} + \frac{1}{12}\left[h^2\frac{\partial^4 U}{\partial x^4}(x_i + \theta_3 h, y_j) + k^2\frac{\partial^4 U}{\partial y^4}(x_i, y_j + \theta_4 k)\right].$$

The differential equation gives

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f$$

and so

$$T_{i,j} = \frac{1}{12}\left[h^2\frac{\partial^4 U}{\partial x^4}(x_i + \theta_3 h, y_j) + k^2\frac{\partial^4 U}{\partial y^4}(x_i, y_j + \theta_4 k)\right].$$

Thus,

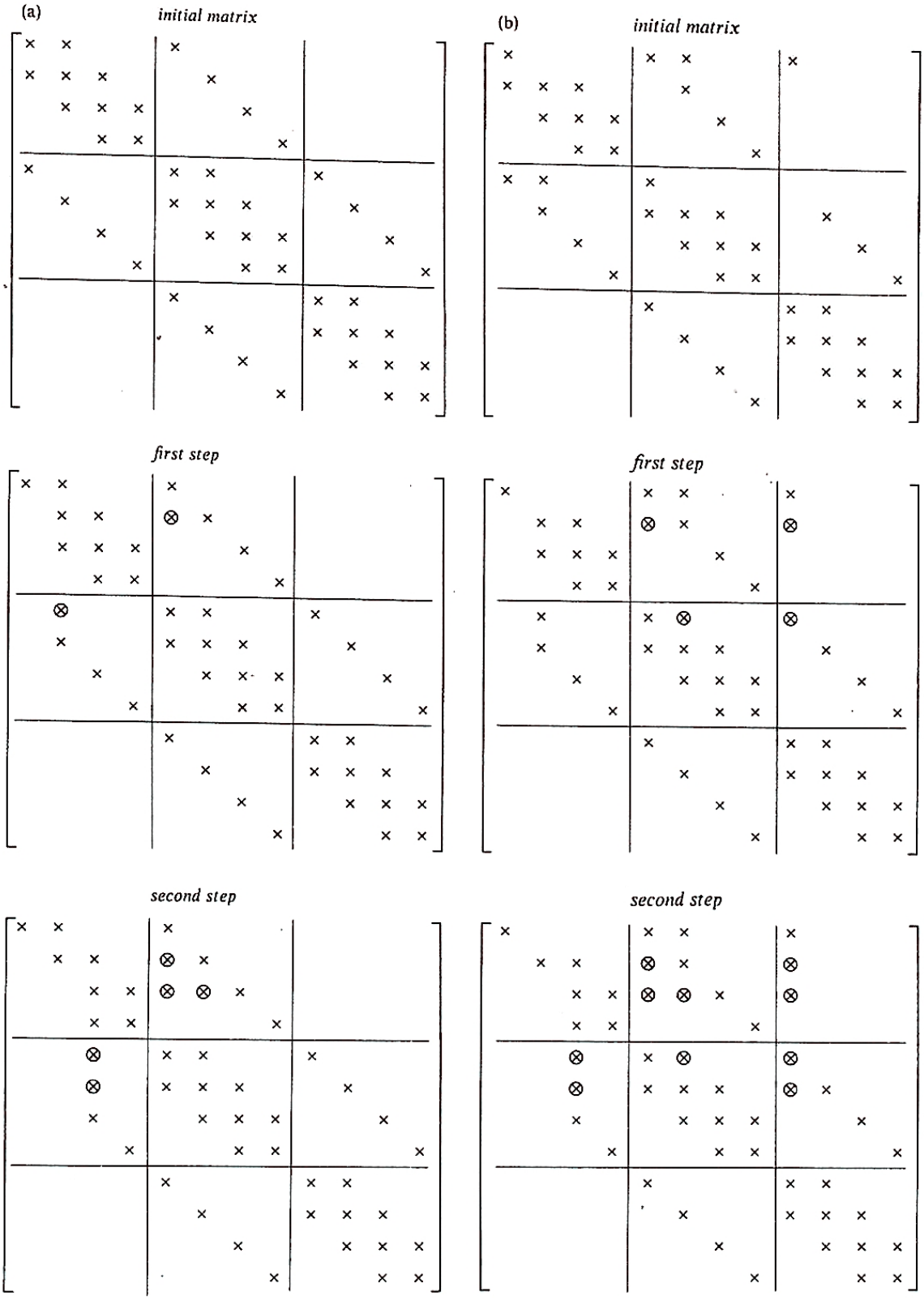
$$|T_{i,j}| \leqslant \frac{1}{12}\left[h^2\left|\frac{\partial^4 U}{\partial x^4}(x_i + \theta_3 h, y_j)\right| + k^2\left|\frac{\partial^4 U}{\partial y^4}(x_i, y_j + \theta_4 k)\right|\right]$$

$$\leqslant \frac{1}{12}(h^2 M_x + k^2 M_y)$$

where M_x and M_y are the largest values of $|\partial^4 U/\partial x^4|$ and $|\partial^4 U/\partial y^4|$ in $D \cup C$, where C is the boundary of D.

Solution to SAQ 5

In the following matrices the entries marked $\times$ are the nonzero entries. The entries marked $\otimes$ are positions of nonzero entries where a zero entry occurred in the initial matrix.



Solution to SAQ 6

For the simple iterative scheme given by

$$\mathbf{x}^{(n)} = (I - A)\mathbf{x}^{(n-1)} + \mathbf{b}$$

we have

$$I - A = \begin{bmatrix} 0.5 & 0.8 \\ 0.2 & 0.5 \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix}.$$

Starting with $\mathbf{x}^{(0)} = (0, 0)$ we obtain

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.5 & 0.8 \\ 0.2 & 0.5 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix},$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0.5 & 0.8 \\ 0.2 & 0.5 \end{bmatrix}\begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix} + \begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 2.15 \\ -0.41 \end{bmatrix}.$$

Similar computations give

$$\mathbf{x}^{(3)} = (2.447, -0.275),$$

$$\mathbf{x}^{(4)} = (2.7035, -0.1481),$$

$$\mathbf{x}^{(5)} = (2.93327, -0.03335).$$

The process should converge to the solution of the linear equations, $\mathbf{x}^* = (5, 1)$ (unique, since $\det A \neq 0$) since the eigenvalues of $I - A$ satisfy

$$(0.5 - \lambda)^2 = 0.16,$$

so that $\lambda_1 = 0.9$, $\lambda_2 = 0.1$. But clearly the convergence is very slow. To verify that $(5, 1)$ is a fixed point of the iteration we check that

$$\begin{bmatrix} 0.5 & 0.8 \\ 0.2 & 0.5 \end{bmatrix}\begin{bmatrix} 5 \\ 1 \end{bmatrix} + \begin{bmatrix} 1.7 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

Solution to SAQ 7

For this simple iterative scheme, the iteration matrix $I - A$ is

$$\begin{bmatrix} 0 & -1 \\ -3 & -6 \end{bmatrix}$$

whose eigenvalues λ are given by $\lambda = -3 \pm \sqrt{12}$. The spectral radius of the iteration matrix is therefore greater than 1 and the scheme will not converge.

Solution to SAQ 8

The iteration matrix of SAQ 6 has eigenvalues given by $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$ (see solution to SAQ 6). Hence the spectral radius $\rho = 0.9$ and the asymptotic rate of convergence of the method is therefore $-\log 0.9 \simeq 0.10$.

Solution to SAQ 9

From the results of SAQ 6 we find

$$\Delta^{(0)} = (1.7, -0.5), \qquad \Delta^{(2)} = (0.297, 0.135),$$

$$\Delta^{(1)} = (0.45, 0.09), \qquad \Delta^{(3)} = (0.2565, 0.1269),$$

$$\Delta^{(4)} = (0.22977, 0.11475).$$

The ratios of the components of $\Delta^{(3)}$ to $\Delta^{(2)}$ are 0.864 and 0.940 and the ratios of the components of $\Delta^{(4)}$ to $\Delta^{(3)}$ are 0.896 and 0.904. We deduce that $\lambda_1 \simeq 0.9$ and expect a better approximation than $\mathbf{x}^{(5)}$ to be

$$\mathbf{x}^{(4)} + \frac{\Delta^{(4)}}{1 - \lambda_1} = \mathbf{x}^{(4)} + 10\Delta^{(4)}$$

$$= (5.0012, 0.9994),$$

which is very close to the true solution!

Solution to SAQ 10

The system can be written as

$$\begin{bmatrix} 1 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 1 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Hence $L = \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$ and $U = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}$.

(a) The Jacobi iteration matrix is

$$J = L + U = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}.$$

(b) We have

$$I - L = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ 0 & -\frac{1}{4} & 1 \end{bmatrix},$$

so that

$$(I - L)^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{16} & \frac{1}{4} & 1 \end{bmatrix}$$

and

$$G = (I - L)^{-1}U = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{16} & \frac{1}{4} \\ 0 & \frac{1}{64} & \frac{1}{16} \end{bmatrix}.$$

(c) For the Jacobi method, $\rho(J) = \max|\alpha|$ where

$$\begin{vmatrix} -\alpha & \frac{1}{4} & 0 \\ \frac{1}{4} & -\alpha & \frac{1}{4} \\ 0 & \frac{1}{4} & -\alpha \end{vmatrix} = 0.$$

Hence

$$-\alpha\left(\alpha^2 - \frac{1}{16}\right) + \frac{1}{16}\alpha = 0$$

and

$$\alpha = 0, \frac{1}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}.$$

Therefore,

$$\rho(J) = \frac{1}{2\sqrt{2}}.$$

For the Gauss-Seidel method, $\rho(G) = \max|\lambda|$ where

$$\begin{vmatrix} -\lambda & \frac{1}{4} & 0 \\ 0 & \frac{1}{16} - \lambda & \frac{1}{4} \\ 0 & \frac{1}{64} & \frac{1}{16} - \lambda \end{vmatrix} = 0.$$

Hence

$$-\lambda^2(\lambda - \tfrac{1}{8}) = 0$$

and

$$\lambda = 0, 0, \tfrac{1}{8}.$$

Therefore,

$$\rho(G) = \tfrac{1}{8}.$$

Clearly $[\rho(J)]^2 = \rho(G)$.

Solution to SAQ11

(a) The given system may be expressed as

$$\begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}.$$

The Jacobi iteration matrix is

$$J = \begin{bmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{bmatrix}$$

and its eigenvalues satisfy

$$\begin{vmatrix} -\lambda & -2 & 2 \\ -1 & -\lambda & -1 \\ -2 & -2 & -\lambda \end{vmatrix} = 0$$

which gives $\lambda^3 = 0$ and all the eigenvalues are zero. The iteration method converges since $\rho(J) = 0$.

The Gauss-Seidel iteration is given by

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} \mathbf{u}^{(n)} = \begin{bmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{u}^{(n-1)} + \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix},$$

which gives

$$\mathbf{u}^{(n)} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{u}^{(n-1)} + \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{bmatrix} \mathbf{u}^{(n-1)} + \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}.$$

The eigenvalues λ of the Gauss-Seidel iteration matrix satisfy

$$\begin{vmatrix} -\lambda & -2 & 2 \\ 0 & 2-\lambda & -3 \\ 0 & 0 & 2-\lambda \end{vmatrix} = 0,$$

that is $\lambda(2 - \lambda)^2 = 0$, and the eigenvalues are 0, 2 and 2. (In fact, this is obvious since the eigenvalues of a triangular matrix are just the elements on the main diagonal.) Since the eigenvalue with largest modulus exceeds unity the method diverges.

(b) The iterative scheme is

$$\mathbf{x}^{(n)} = \begin{bmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{bmatrix} \mathbf{x}^{(n-1)} + \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}.$$

Taking $\mathbf{x}^{(0)} = (\alpha, \beta, \gamma)$, we obtain

$$\mathbf{x}^{(1)} = (-2\beta + 2\gamma + 1, -\alpha - \gamma + 3, -2\alpha - 2\beta + 5),$$

$$\mathbf{x}^{(2)} = (-2\alpha - 4\beta + 2\gamma + 5, 2\alpha + 4\beta - 2\gamma - 3, 2\alpha + 4\beta - 2\gamma - 3)$$

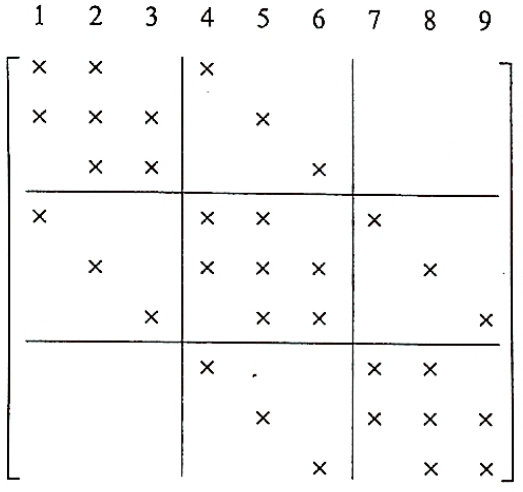
and $\mathbf{x}^{(3)} = (1, 1, 1)$.

The third iterate is independent of (α, β, γ), and $\mathbf{x}^{(3)}$ is the correct solution of the linear equations.

This result is to be expected, since $\rho(J) = 0$ so that the correction vector is eventually zero, i.e. the scheme terminates after a finite number of steps.

Solution to SAQ 12

The form of the matrix for the natural ordering is



and the partitioning shows that it is not the shape of either of the required forms, since the diagonal blocks are not of diagonal form.

Solution to SAQ 13

(a) The statements are equivalent because the SOR method merely changes the Gauss-Seidel adjustments by a constant factor ω. Alternatively, the SOR iteration is given by

$$(I - \omega L)\mathbf{u}^{(n+1)} = [(1 - \omega)I + \omega U]\mathbf{u}^{(n)} + \omega\mathbf{b}$$

and the Gauss-Seidel by

$$(I - L)\mathbf{u}^{(n+1)} = U\mathbf{u}^{(n)} + \mathbf{b}.$$

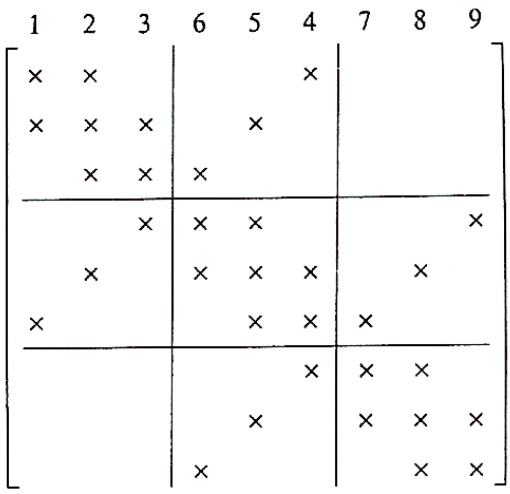
In each case the points with superscript $(n + 1)$ are related to the matrix on the left of the equation, and the nonzero elements in $I - \omega L$ are in the same positions as those in $I - L$.

(b) For the natural ordering we have the following table (corresponding to the table in the text relevant to the diagonal ordering):

	points with superscript $(n + 1)$	*points with superscript* (n)
$r = 1$	none	2, 4
2	1	5, 3
3	2	6
4	1	7, 5
5	4, 2	8, 6
6	5, 3	9
7	4	8
8	7, 5	9
9	8, 6	none

This is the same as the table in the text. Therefore, the natural ordering is consistent with the diagonal ordering of Equation (6).

(c) The matrix for the ordering 1, 2, 3; 6, 5, 4; 7, 8, 9 is given by



If we look at the computation of $u_6^{(n+1)}$ (i.e., the fourth row) we see that it involves $u_3^{(n+1)}$, $u_5^{(n)}$ and $u_9^{(n)}$. This is not consistent with $r = 6$ in the table in the text for the diagonal scheme, which involves $u_5^{(n+1)}$. In the checker board scheme we find that $u_6^{(n+1)}$ is computed from a knowledge of $u_3^{(n+1)}$, $u_5^{(n+1)}$ and $u_9^{(n+1)}$, so that there is no consistency here either.

Solution to SAQ 14

(a) For the problem of SAQ 10 we found $[\rho(J)]^2 = \frac{1}{8}$. The optimum ω for the SOR method is given by

$$\omega_{opt} = \frac{2}{1 + (1 - \frac{1}{8})^{\frac{1}{2}}} \simeq 1.039.$$

The spectral radius of the SOR matrix is then $\omega_{opt} - 1 \simeq 0.039$, considerably smaller than the 0.125 of the Gauss-Seidel method.

The matrix of the equations has Property *A* since it has the form of Equation (8) of Section 11.4.3; all the block matrices are 1×1 matrices. It is also consistent since no interchanges are required to transform the matrix to this form.

(b) The following is a sample computer output from the library program $SOR321 in which both the Gauss-Seidel method (over-relaxation factor = 1.0) and SOR method (over-relaxation factor = 1.039) have been used to solve the given system starting with the initial approximation $(u_1, u_2, u_3) = (0, 0, 0)$.

```
THIS PROGRAM USES THE METHOD OF SUCCESSIVE OVER RELAXATION
TO SOLVE A SYSTEM OF LINEAR EQUATIONS TO SIX DEC.PL.

TYPE NUMBER OF EQUATIONS(<=10)?3

TYPE ELEMENTS OF COEFFICIENT MATRIX,ROW BY ROW
ROW 1     ?1,0.25,0
ROW 2     ?0.25,1,0.25
ROW 3     ?0,0.25,1

TYPE R.H.S. CONSTANTS
?1.5,3,3.5

TYPE INITIAL APPROXIMATION
?0,0,0

TYPE OVER-RELAXATION FACTOR?1.0
```

```
SOLUTION
--------
 1

 2

 3


NUMBER OF  ITERATIONS NEEDED IS 10

REPEAT WITH ANOTHER OVER-RELAXATION FACTOR/INITIAL APPROX.
YES OR NO?YES
TYPE INITIAL APPROXIMATION
?0,0,0

TYPE OVER-RELAXATION FACTOR?1.039


SOLUTION
--------
 1

 2

 3


NUMBER OF ITERATIONS NEEDED IS 7

REPEAT WITH ANOTHER OVER-RELAXATION FACTOR/INITIAL APPROX.
YES OR NO?NO
```

11.7 APPENDIX (*Optional*)

The Eigenvalues of a Block Tridiagonal Matrix

We show how to derive the eigenvalues of the block tridiagonal matrix A of order $(p-1)(q-1)$ given by

$$A = \begin{bmatrix} B & I & & & & \\ I & B & I & & & \\ & I & B & I & & \\ & & \cdot & \cdot & \cdot & \\ & & & I & B & I \\ & & & & I & B \end{bmatrix},$$

where B is a square matrix of order $q-1$ and I is the identity matrix of order $q-1$. If λ is an eigenvalue of A and $\mathbf{x} = (x_{11}, x_{12}, \ldots, x_{1,q-1}; x_{21}, \ldots, x_{p-1,q-1})$ is a corresponding eigenvector then

$$\begin{bmatrix} B-\lambda I & I & & & & \\ I & B-\lambda I & I & & & \\ & I & B-\lambda I & I & & \\ & & \cdot & \cdot & \cdot & \\ & & & I & B-\lambda I & I \\ & & & & I & B-\lambda I \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \cdot \\ \cdot \\ \mathbf{x}_{p-2} \\ \mathbf{x}_{p-1} \end{bmatrix} = \mathbf{0},$$

where $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{k,q-1})$. This yields the vector recurrence relation

$$\mathbf{x}_{k+1} + (B - \lambda I)\mathbf{x}_k + \mathbf{x}_{k-1} = \mathbf{0}$$

together with the boundary conditions

$$\mathbf{x}_0 = \mathbf{x}_p = \mathbf{0}.$$

We now write $\nu = \lambda - \mu$, where μ is an eigenvalue of B with a corresponding eigenvector $\mathbf{b}$ and suppose that

$$\mathbf{x}_k = \alpha_k \mathbf{b}.$$

Hence

$$(\alpha_{k+1} - \nu\alpha_k + \alpha_{k-1})\mathbf{b} = \mathbf{0}$$

and, since $\mathbf{b} \neq \mathbf{0}$ by the definition of an eigenvector, we have

$$\alpha_{k+1} - \nu\alpha_k + \alpha_{k-1} = 0,$$

$$\alpha_0 = \alpha_p = 0,$$

or equivalently,

$$\begin{bmatrix} -\nu & 1 & & & & \\ 1 & -\nu & 1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & 1 & -\nu & 1 \\ & & & & 1 & -\nu \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_{p-2} \\ \alpha_{p-1} \end{bmatrix} = 0.$$

That is, ν is an eigenvalue of the square matrix of order $p-1$

$$\begin{bmatrix} 0 & 1 & & & & & \\ 1 & 0 & 1 & & & & \\ & 1 & 0 & 1 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & 1 & 0 & 1 \\ & & & & & 1 & 0 \end{bmatrix}$$

and is given by

$$\nu_i = 2\cos\frac{i\pi}{p} \qquad i = 1, 2, \ldots, p-1$$

with an associated eigenvector given by

$$\left(\sin\frac{i\pi}{p}, \sin\frac{2i\pi}{p}, \ldots, \sin\frac{(p-1)i\pi}{p}\right),$$

as shown in the Appendix to *Unit 8*.

Hence we obtain the result that the eigenvalues of the block tridiagonal matrix A are given by

$$\lambda_{i,j} = \mu_j + \nu_i \qquad i = 1, 2, \ldots, p-1, \qquad j = 1, 2, \ldots, q-1,$$

where μ_j are the eigenvalues of B and

$$\nu_i = 2\cos\frac{i\pi}{p} \qquad i = 1, 2, \ldots, p-1.$$

The corresponding eigenvector is given by

$$\mathbf{x}^{(i,j)} = \left(\sin\frac{i\pi}{p}\mathbf{b}_j; \sin\frac{2i\pi}{p}\mathbf{b}_j; \cdots; \sin\frac{(p-1)i\pi}{p}\mathbf{b}_j\right).$$

Unit 13 Sturm–Liouville Theory

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differential Equations* (Oxford, 1971).

H. F. Weinberger, *A First Course in Partial Differential Equations* (Xerox, 1965).

It is essential to have these books; the course is based on them and will not make sense without them. They are referred to in the text as *S* and *W* respectively.

Unit 13 is based on *W*: Chapter VII, Sections 36 to 38.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics.* References to Open University courses in mathematics take the form:

> *Unit M100 13, Integration II* for the Mathematics Foundation Course,
> *Unit M201 23, The Wave Equation* for the Linear Mathematics Course.

13.0 INTRODUCTION

This unit is concerned with properties of the eigenfunctions and eigenvalues of the system

$$\begin{aligned} &(pu')' - qu + \lambda \rho u = 0 \qquad \text{in } (\alpha, \beta), \\ &u(\alpha) = u(\beta) = 0, \end{aligned} \tag{1}$$

and of similar systems having different, homogeneous, boundary conditions. Such systems are known as *Sturm–Liouville systems* and arise, for example, when the separation of variables technique is applied to solve boundary value problems for partial differential equations; so far in this course we have only looked at the simplest types, where p, q and ρ are constant functions.

The differential equation (1) may have non-constant functions as coefficients and we may not be able to solve it explicitly using analytical methods. However, the art of dealing with difficult problems is to discover useful things about the solution without necessarily being able to express it as a formula. The present unit is a collection of results of this sort. Illustrations are chosen for which solutions *can* be written down, but the results apply equally to systems about which we would otherwise be wholly in the dark.

The results can be summed up very briefly: the system (1) has most of the properties of the simplest of all such systems, namely

$$\begin{aligned} &u'' + \lambda u = 0 \qquad \text{in } (0, 1), \\ &u(0) = u(1) = 0. \end{aligned} \tag{2}$$

This system has eigenvalues $\lambda_k = k^2\pi^2$, and eigenfunctions $u_k(x) = A_k \sin k\pi x$. The eigenvalues are discrete, positive and tend to infinity, and the eigenfunctions *oscillate*, u_k having $k - 1$ zeros in $(0, 1)$. The eigenfunctions can be used in infinite series (Fourier sine series) to represent more or less arbitrary functions on $[0, 1]$, and for sufficiently well-behaved functions the series converge uniformly on $[0, 1]$. These properties are all shared with the solutions of the general system (1). It can even be shown that, for k large, the eigenvalues of (1) are nearly proportional to those of (2). Practical instances of this theory appear in *Unit 14, Bessel Functions.*

In order to prove these results, the system (1) will be reformulated in two different ways; as an *integral equation* in which the Green's function is used, and as a *minimization problem.* The second idea will be completely new to you and you should take a lot of trouble to make sure you understand it.

Some of the SAQs have been marked as being difficult. Do not spend a lot of time trying to do them yourself, and if in difficulty go straight to the answer.

13.1 EIGENVALUES AND EIGENFUNCTIONS

13.1.1 Elementary Properties

READ W: page 160 to page 162, line −6.

Notes

(i) *W: page 160, lines 1 to 4*
See, for example, *W: Section 14.*

(ii) *W: page 162, Equation (36.7)*
The system (36.1), (36.2) may be written as

$$(pu')' - qu = -\lambda\rho u \qquad \text{in } (0, 1), \tag{1}$$
$$u(0) = u(1) = 0.$$

Suppose the system has a solution, that is, λ is an eigenvalue. Since all the eigenvalues are positive (*W: page 161, line −11*), $\lambda = 0$ is not an eigenvalue and the system

$$(pu')' - qu = 0 \qquad \text{in } (0, 1), \tag{2}$$
$$u(0) = u(1) = 0$$

has only the trivial solution $u = 0$. Therefore, a Green's function $G(x, \xi)$ exists for this system (*Unit 9, Green's Functions I*, SAQ 9). Now, for the system (36.6) the equation in *W: page 162, line 2* holds for every function F. It therefore holds when

$$F(x) = \lambda\rho(x)u(x),$$

λ being an eigenvalue and u a corresponding eigenfunction of (1). Thus we have (36.7).

Equation (36.7) is an *integral equation*, which means that the unknown function occurs under an integral sign. Neither u nor λ is known in this equation. Note that the boundary conditions are automatically incorporated in (36.7), since the integral tends to zero as $x \rightsquigarrow 0$ or $x \rightsquigarrow 1$.

In the subsequent work, you should notice that we do not have to know what G is, but only that there exists a Green's function belonging to the system (2) and that it has some simple properties.

(iii) *W: page 162, line −10*
We met the result that

$$\sum_1^\infty a_n \text{ converges} \Rightarrow \lim_{n \rightsquigarrow \infty} a_n = 0$$

in *Unit M201 20, Convergence and Bases.* Here we put $a_n = 1/\lambda_n^2$, and deduce that λ_n^2 grows unboundedly. Since $\lambda_n > 0$, we have $\lambda_n \longrightarrow \infty$.

It is assumed that the eigenvalues of the given system form a discrete set of numbers. This will follow from the *minimum principle* described in Section 13.1.2.

SAQ 1

Find the eigenvalues and eigenfunctions of the system

$$u'' + \lambda u = 0 \qquad \text{in } (\alpha, \beta),$$
$$u(\alpha) = u(\beta) = 0.$$

(Solution on p. 28.)

SAQ 2

Find the eigenvalues and eigenfunctions of the system

$$u'' + \lambda u = 0 \qquad \text{in } (\alpha, \beta),$$
$$u'(\alpha) = u'(\beta) = 0.$$

In *W: page 161* it was proved that the system (36.1), (36.2) has no eigenvalue $\lambda = 0$, but $\lambda = 0$ is an eigenvalue of the system above. Where does the theory need changing for this system?

(Solution on p. 28.)

SAQ 3

The system

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (\alpha, \beta),$$

$$u(\alpha) = 0,\ u(\beta) = 0,$$

has an eigenvalue λ_k and a corresponding eigenfunction u_k. Show that every eigenfunction of λ_k is a multiple of u_k, that is to say, the dimension of the eigenspace is one.

HINT: Suppose there are two linearly independent eigenfunctions. Then every solution of the equation with $\lambda = \lambda_k$ must be a linear combination of these. This leads to a contradiction.

(Solution on p. 29.)

SAQ 4

For the following system, prove that the eigenfunctions corresponding to different eigenvalues are orthogonal with respect to the weight function ρ.

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (\alpha, \beta),$$

$$Au'(\alpha) + Bu(\alpha) = Cu'(\beta) + Du(\beta) = 0,$$

where A and C are nonzero.

(Solution on p. 30.)

SAQ 5

Find the eigenvalues and eigenfunctions of the system

$$\frac{d^2u}{d\theta^2} + \lambda u = 0 \qquad -\pi < \theta < \pi,$$

with *periodic* boundary conditions,

$$u(-\pi) = u(\pi),\ u'(-\pi) = u'(\pi).$$

Show that the solution space corresponding to each nonzero eigenvalue has dimension 2.

(Solution on p. 30.)

SAQ 6

Given the system

$$u'' + \lambda u = 0 \qquad \text{in } (0, 2\pi),$$

$$u(0) = 0,$$

$$u'(2\pi) - u(2\pi) = 0,$$

show that the eigenvalues are the solutions λ of the equation

$$\lambda^{\frac{1}{2}} = \tan(2\pi\lambda^{\frac{1}{2}}),$$

and that the least eigenvalue is negative.

(Solution on p. 31.)

13.1.2 The Minimum Principle

In Section 13.1.1 we succeeded in obtaining certain properties of the eigenvalues by replacing the differential equation system (36.1), (36.2) in *W* by an integral equation. In this section we reformulate the problem in another way which may be completely new to you. It is a method of wide application, and represents a special case in the subject called the *calculus of variations*. The simplest problem in this subject is to find a function which minimizes the numerical value of a given definite integral in which the function appears. For example, the function ϕ minimizing

$$\int_0^{\pi/2} [\phi^2(x) - \phi'^2(x)]\, dx$$

over functions satisfying $\phi(0) = 0$, $\phi(\frac{1}{2}\pi) = 1$ is given by $\phi(x) = \sin x$.

READ W: page 162, line −5 to page 165, line 9.

Notes

(i) *W: page 163, line 6*
The passage leading up to (36.8) is intended merely to motivate an approach which *begins* with (36.8), since suddenly to begin examining the form (36.8) without any suggestion that it might be profitable to do so would be unreasonable. There is therefore no need to learn this argument.

(ii) *W: page 163, lines 8 to 10*
This is a use of Parseval's equation in its general form, *W: page 75, Equation (16.6)*, with

$$\frac{1}{\rho}[(p\phi')' - q\phi]$$

in place of f^*.

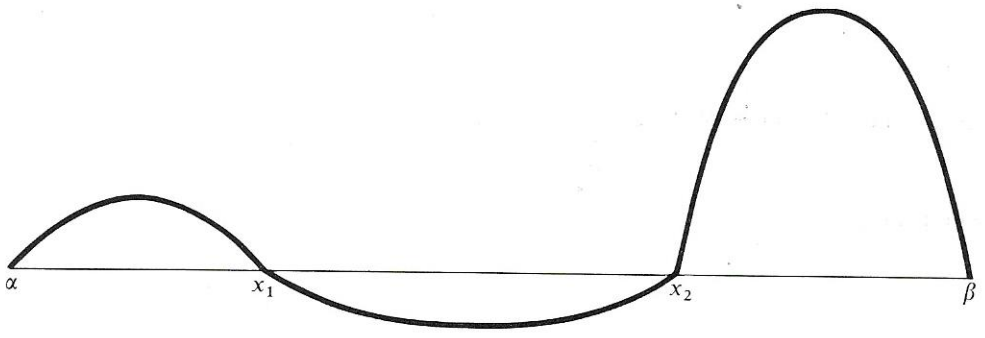
Since $\lambda_n \geqslant \lambda_1$ for $n = 1, 2, 3, \ldots$, we now have

$$\sum_{n=1}^{\infty} \lambda_n c_n^2 \int_0^1 \rho u_n^2 \geqslant \lambda_1 \sum_{n=1}^{\infty} c_n^2 \int_0^1 \rho u_n^2 = \lambda_1 \int_0^1 \rho \phi^2,$$

by the equation on *line 7*.

(iii) *W: page 163, line −12*

A function is said to be **piecewise continuously differentiable** or **piecewise smooth** on a finite interval $[\alpha, \beta]$ if the interval can be divided into a finite number of open subintervals $(\alpha, x_1), (x_1, x_2), \ldots, (x_{n-1}, \beta)$, such that the function and its first derivative are continuous on each subinterval and the right- and left-hand limits of the function and its first derivative exist at the end points of each subinterval. This class of functions embraces most functions directly encountered in physical applications. The following figure shows a continuous, piecewise continuously differentiable function.



(iv) *W: page 163, line −10*

For example, put $p \equiv 1$, $q \equiv 0$, $\rho \equiv 1$. The system becomes

$$u'' + \lambda u = 0 \qquad \text{in } (0, 1), \qquad u(0) = u(1) = 0,$$

with eigenvalues $\lambda = n^2\pi^2$, the first eigenvalue λ_1 being equal to π^2. Then what is suggested (not proved) in the preceding paragraph is that for any continuous, piecewise continuously differentiable function ϕ satisfying $\phi(0) = \phi(1) = 0$,

$$\frac{\int_0^1 \phi'^2}{\int_0^1 \phi^2} \geqslant \pi^2.$$

For example, if $\phi(x) = x(1 - x)$ we obtain for the left-hand side

$$\frac{\int_0^1 (1 - 4x + 4x^2)\,dx}{\int_0^1 (x^2 - 2x^3 + x^4)\,dx} = \frac{\frac{1}{3}}{\frac{1}{30}} = 10$$

which is greater than $\pi^2 \simeq 9.87$ by a small amount. You may try other functions, but will always fail to produce a value less than π^2, although the function $\phi(x) = \sin x$ gives equality. Therefore the result seems likely to be true.

So, instead of trying to refine the given proof so that the gap in *line 6* is filled, we restart from the standpoint of Equation (36.8), by proving directly that if there exists an admissible function ϕ which minimizes the value of this expression, then the minimum value attained is the lowest eigenvalue, and the function ϕ is a corresponding eigenfunction.

The proof that there is a minimum, given by some admissible function, is beyond the scope of this course. Generally speaking, to prove the existence of something without being able to display it is a difficult matter. It is easy to show similar expressions that have no minimum: for instance,

$$\frac{\int_0^1 \phi'^2}{\int_0^1 \phi}$$

has no minimum. We shall simply assume throughout this unit that there does exist an admissible minimizing function for the Rayleigh quotient.

(v) *W: page 163, line −7*

ϕ is admissible if ϕ is continuous on [0, 1], ϕ' is piecewise continuous on [0, 1], and $\phi(0) = \phi(1) = 0$.

(vi) *W: page 163, lines −5 to −3*

ψ gives the minimum value, μ, to the Rayleigh quotient, and therefore for any other admissible function, such as $\psi + \tau\phi$, the "$\geqslant$" in *line −5* holds. (Note that ϕ, though still an arbitrary admissible function, has a different significance from its earlier use.) In particular, for any given admissible ϕ, the "$\geqslant$" holds for every value of τ, and becomes an equality when $\tau = 0$. Therefore, if ϕ is a given function, the left-hand side is a smooth function of τ with a minimum at $\tau = 0$.

(vii) *W: page 164, lines 2 to 4*

The integration by parts gives

$$-\int_0^1 \phi[(p\psi')' - q\psi + \mu\rho\psi] + \Big[p\psi'\phi\Big]_0^1 = 0.$$

The bracket vanishes because $\phi(0) = \phi(1) = 0$; this is the stage in the argument where the boundary conditions appear. Note that we are now assuming that a minimizing function ψ not only exists, but also is twice continuously differentiable throughout (0, 1).

(viii) *W: page 164, last paragraph*

Let $\phi = \psi$ be that function which minimizes the Rayleigh quotient amongst all continuous, piecewise continuously differentiable functions ϕ satisfying

$$\phi(0) = \phi(1) = 0$$

and

$$\int_0^1 \rho\phi u_1 = 0. \tag{1}$$

Now the proof proceeds exactly as before up to *W: page 164, line 12*, i.e.

$$(p\psi')' - q\psi + \mu\rho\psi = 0,$$

$$\psi(0) = \psi(1) = 0.$$

ψ is therefore an eigenfunction which is, by (1), orthogonal to u_1. It must therefore be an eigenfunction with an eigenvalue different from λ_1 (by SAQ 3) and the smallest available eigenvalue is λ_2, with eigenfunction u_2.

The proof for subsequent eigenvalues is treated in the same way.

(ix) *W: page 165, lines 4 to 7*

This shows that, for any k, the family of admissible functions which are orthogonal to the first $k - 1$ eigenfunctions is nonempty—a polynomial is constructed. That is to say, no matter how large k may be, we have admissible functions for the next step.

SAQ 7

Verify by substitution that

$$\frac{\int_0^1 u_1'^2}{\int_0^1 u_1^2} = \lambda_1,$$

where λ_1 is the lowest eigenvalue and u_1 any corresponding eigenfunction of the system

$$u'' + \lambda u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u(1) = 0.$$

(Solution on p. 31.)

SAQ 8

Show that the smallest eigenvalue of the system

$$u'' + \lambda u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u'(1) - u(1) = 0,$$

is arrived at by minimizing

$$R[\phi] = \frac{\int_0^1 \phi'^2 - [\phi(1)]^2}{\int_0^1 \phi^2}$$

over all continuous piecewise continuously differentiable functions ϕ satisfying $\phi(0) = \phi'(1) - \phi(1) = 0$. You may assume that a twice continuously differentiable minimizing function exists.

(Solution on p. 32.)

SAQ 9

Let λ_k be the kth eigenvalue and u_k the kth eigenfunction of the system with periodic boundary conditions,

$$u'' + \lambda u = 0 \qquad \text{in } (-\pi, \pi)$$

$$u(-\pi) = u(\pi),\ u'(-\pi) = u'(\pi),$$

which you solved in SAQ 5. (Double eigenvalues are to be counted twice.)

Proceed as in *W: pages 163 to 165* to show that

$$\lambda_k = \min \frac{\int_{-\pi}^{\pi} \phi'^2}{\int_{-\pi}^{\pi} \phi^2}$$

the minimum being taken over functions satisfying

$$\phi(-\pi) = \phi(\pi),\ \phi'(-\pi) = \phi'(\pi),$$

and subject to the orthogonality conditions

$$\int_{-\pi}^{\pi} \phi u_1 = \cdots = \int_{-\pi}^{\pi} \phi u_{k-1} = 0.$$

(Solution on p. 32.)

SAQ 10 (*Very difficult*)

Prove that the minimum of the quotient

$$\frac{\int_0^1 (p\phi'^2 + q\phi^2)}{\int_0^1 \rho\phi^2},$$

over all smooth functions ϕ satisfying only the condition $\phi(0) = 0$, is equal to the lowest eigenvalue of the system

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u'(1) = 0,$$

and the minimizing function is a corresponding eigenfunction. By considering a particular choice of p, q and ρ deduce that

$$\int_0^1 \phi'^2 \geqslant \tfrac{1}{4}\pi^2 \int_0^1 \phi^2,$$

for every smooth ϕ for which $\phi(0) = 0$.

(Solution on p. 33.)

SAQ 11

A certain wave propagation problem inside a sphere of radius a gives rise to the following system:

$$\frac{d^2u}{dr^2} + \frac{2}{r}\frac{du}{dr} + \lambda u = 0 \qquad 0 < r < a,$$

$$u(a) = 0,\ u(0) \text{ is finite,}$$

where r is the distance from the centre.

(a) Find all the solutions.

HINT: Determine the differential equation satisfied by ru.

(b) Suggest a minimization problem equivalent to this one. (Remember to use the self-adjoint form.)

(Solution on p. 34.)

SAQ 12 (*Difficult*)

Let p, q and ρ have the same meanings as in *W*, and let v, w be any two piecewise continuously differentiable functions on [0, 1]. Now define E and H by

$$E(v, w) = \int_0^1 (pv'w' + qvw),$$

and

$$H(v, w) = \int_0^1 \rho vw.$$

(a) Write the Rayleigh quotient in terms of E and H.

(b) Prove that H defines an inner product.*

(c) Prove that E defines an inner product when $q(x) > 0$ for all $x \in (0, 1)$.

(d) Carry out the proof that the minimum of the Rayleigh quotient is equal to the lowest eigenvalue (*W: page 164, Equation (36.10)*) using this notation.

(e) Show that any function solving Problem B also solves Problem A:

Problem A

Minimize

$$\frac{\int_0^1 (p\phi'^2 + q\phi^2)}{\int_0^1 \rho\phi^2}$$

over all smooth functions for which $\phi(0) = \phi(1) = 0$.

Problem B

Minimize

$$\int_0^1 (p\phi'^2 + q\phi^2)$$

over all smooth functions for which $\phi(0) = \phi(1) = 0$, and

$$\int_0^1 \rho\phi^2 = 1.$$

(Solution on p. 34.)

*The definition of an inner product is given in the Appendix to *Unit 3, Elliptic and Parabolic Equations.*

13.2 THE CONVERGENCE OF GENERAL FOURIER SERIES

13.2.1 The General Theory

The detailed analysis involved in this section is very difficult, and we therefore recommend that you simply *read through it, noticing the principal results.* Some of the details are not spelled out in *W*, and in Section 13.2.2 we prove these results. If you are pressed for time you should omit this section altogether and continue with Section 13.2.3. The general nature of the results is as follows. Let $\{u_n\}$ be a sequence of eigenfunctions of the system (36.1), (36.2) in *W: page 160.* It has been proved in Section 13.1 that they form an infinite set of orthogonal functions. We have seen in *Unit 6, Fourier Series* that it is possible to expand an arbitrary function f on the interval (0, 1) in terms of these eigenfunctions. We should like to be able to say further that if

$$c_n = \frac{\int_0^1 \rho(x) f(x) u_n(x)\, dx}{\int_0^1 \rho(x) u_n^2(x)\, dx}$$

then

$$f(x) = \sum_1^\infty c_n u_n(x) \qquad 0 \leqslant x \leqslant 1,$$

and that the convergence is uniform (this being essentially the result for trigonometric Fourier series provided f is suitably restricted).

In the next reading passage it is shown that the eigenfunctions $\{u_n\}$ are as good as the trigonometric functions for this purpose. In fact the convergence properties of trigonometric Fourier series follow as a particular case of the more general series above.

READ W: page 165, line 10 to page 168, line −12 (the end of the section).

Notes

(i) *W: page 165, line 16*
The functions u_i and u_j are orthogonal when $i \neq j$.

(ii) *W: page 165, Equation (36.12)*
Note that, for example,

$$\left(\sum_1^{k-1} c_n u_n\right)^2 = \sum_{n=1}^{k-1} c_n u_n \sum_{m=1}^{k-1} c_m u_m$$
$$= \sum_{n=1}^{k-1} \sum_{m=1}^{k-1} c_n c_m u_n u_m.$$

(iii) *W: page 166, line 2*
Since the terms in the series are positive the right-hand side is bounded above by

$$\frac{1}{\lambda_k} \int_0^1 (pf'^2 + qf^2),$$

which approaches zero as $\lambda_k \rightsquigarrow \infty$.

(iv) *W: page 166, lines 6 to 8*
The proof that any square integrable function can be approximated in the mean arbitrarily closely by a continuously differentiable function is beyond the scope of this course. (However, if you have ample time at your disposal, you may like to sketch out a proof that a *piecewise smooth* function can be approximated in the mean arbitrarily closely by a continuously differentiable function.) The proof of completeness is given as Theorem A in Section 13.2.2.

(v) *W: page 166, lines 9 and 10*
This statement means that if

$$c_n = \frac{\int_0^1 \rho f u_n}{\int_0^1 \rho u_n^2} \qquad n = 1, 2, 3, \ldots$$

then

$$\lim_{k \to \infty} \int_0^1 \rho \left(f - \sum_{r=1}^{k} c_r u_r \right)^2 = 0.$$

(vi) *W: page 167, lines 1 to 3*
From *W: page 166, line −4*, with $f(\xi) = G(x, \xi)$, we have

$$G(x, \xi) \sim \sum_1^\infty c_n u_n(\xi),$$

where the Fourier coefficients $\{c_n\}$ depend on x. They are given by

$$c_n = \frac{\int_0^1 \rho(\xi) G(x, \xi) u_n(\xi)\, d\xi}{\int_0^1 \rho(\xi)[u_n(\xi)]^2\, d\xi}$$

according to the general formula at the bottom of *W: page 162.*

But by Equation (36.7), with $\lambda = \lambda_n$ and $u = u_n$,

$$\int_0^1 \rho(\xi) G(x, \xi) u_n(\xi)\, d\xi = \frac{1}{\lambda_n} u_n(x),$$

and therefore

$$c_n = \frac{1}{\lambda_n} \frac{u_n(x)}{\int_0^1 \rho(\xi)[u_n(\xi)]^2\, d\xi}$$

as in *line 2*; d_n is treated similarly.

(vii) *W: page 167, Equation (36.14)*
Here we have expressed the Green's function for the system (36.6) in terms of the eigenvalues and eigenfunctions of an associated eigenvalue problem (36.1), (36.2), in which ρ is any positive continuous function of our choice with domain [0, 1]. This formula is called the **bilinear formula** for the Green's function. Note that the symmetry of G is clearly displayed in the series.

(viii) *W: page 167, line −9*
Schwarz's Inequality in a finite dimensional Euclidean space gives

$$\left(\sum_{i=1}^{k} a_i b_i \right)^2 \leqslant \left(\sum_{i=1}^{k} a_i^2 \right) \left(\sum_{i=1}^{k} b_i^2 \right),$$

for any real numbers $a_1, \ldots, a_k, b_1, \ldots, b_k$.

General Comment

From SAQ 9 we see that the trigonometric functions,

$$\cos n\theta, \sin n\theta \qquad n = 1, 2, \ldots \tag{1}$$

are the eigenfunctions of a system of the type considered, but with periodic boundary conditions, on the interval $[-\pi, \pi]$. They also satisfy a minimization principle similar to (36.11), with the appropriate modification to the boundary conditions. By making the small changes in the theory arising from the changes in the interval and boundary conditions we can show that the set of functions (1) is complete on $[-\pi, \pi]$, and also deduce the standard theorems on convergence which we quoted without proof in *Unit 6.*

SAQ 13

The Green's function for the problem

$$u''(x) = -f(x) \qquad 0 < x < 1,$$

is given by

$$G(x, \xi) = \begin{cases} (1 - \xi)x & x \leqslant \xi \\ \xi(1 - x) & x \geqslant \xi \end{cases}$$

(Example, *W: page 122*). Use the bilinear formula (36.14) to express G as a Fourier series on $[0, 1]$.

HINT: Make a suitable choice for ρ.

(Solution on p. 35.)

SAQ 14

Prove that the set of functions $\sin n\pi x$ $(n = 1, 2, \ldots)$ is complete on $0 \leqslant x \leqslant 1$. (This is essentially the set used for expansions in sine series; see *W: Section 20*. The convergence properties used there follow from the results of this section.)

(Solution on p. 36.)

13.2.2 Three Theorems (Optional)

We now fill in the details of three main results which were stated in *W* with only brief proofs. *You should not spend time on this section at the expense of the remainder of the unit.*

The three results are:

A the eigenfunctions $\{u_n\}$ of the problem (36.1), (36.2) are complete (*W: page 166*);

B the Fourier series of a continuous function f with $f(0) = f(1) = 0$ and

$$\int_0^1 (pf'^2 + qf^2)$$

finite, in terms of the complete set of eigenfunctions $\{u_n\}$, converges uniformly to f on $[0, 1]$ (*W: page 167*);

C the solution of the nonhomogeneous problem is given by the formula in *W: page 168, line 5*.

We first quote the *Triangle Inequality** for the space of functions square integrable on $(0, 1)$ with weight function ρ; this tells us that for any two such functions g and h we have

$$\left[\int_0^1 \rho(g + h)^2\right]^{\frac{1}{2}} \leqslant \left[\int_0^1 \rho g^2\right]^{\frac{1}{2}} + \left[\int_0^1 \rho h^2\right]^{\frac{1}{2}}.$$

This result will be used in the proof of Theorem A. The essential point to be proved in this theorem is that a function f can be approximated in the mean by its Fourier series in terms of the eigenfunctions $\{u_n\}$. It has already been shown in *W: page 166, line 4* that if f is continuously differentiable then this result holds. We wish to extend the result to more general functions which arise and can themselves be approximated in the mean by continuously differentiable functions.

*The Triangle Inequality may be deduced from Schwarz's Inequality in any inner product space: for example, D. L. Kreider *et al.*, *An Introduction to Linear Analysis* (Addison-Wesley, 1966), p. 264.

THEOREM A

For all f such that $\int_0^1 \rho f^2$ is finite,

$$\lim_{k \rightsquigarrow \infty} \int_0^1 \rho \left(f - \sum_1^k c_n u_n \right)^2 = 0,$$

where $\{c_n\}$ is the sequence of Fourier coefficients of f with respect to $\{u_n\}$.

Proof

Let f be any function of the class considered. Then given any $\varepsilon > 0$, let $\bar{f}$ be a continuously differentiable function such that

$$\int_0^1 \rho(f - \bar{f})^2 < \varepsilon \quad \text{and} \quad \bar{f}(0) = \bar{f}(1) = 0. \tag{1}$$

(As stated in note (iv) of Section 13.2.1 the proof that such an $\bar{f}$ exists is omitted from the course.) Now let $\{d_n\}$ be the sequence of Fourier coefficients for $\bar{f}$, and put

$$s_k(x) = \sum_1^k c_n u_n(x), \qquad \sigma_k(x) = \sum_1^k d_n u_n(x),$$

for the kth partial sums. For any given k the integral

$$\int_0^1 \rho \left(f - \sum_1^k b_n u_n \right)^2$$

is minimized when the sequence $\{b_n\}$ is the sequence of Fourier coefficients (*W: page 71*). Therefore, for every k,

$$0 \leqslant \int_0^1 \rho(f - s_k)^2 \leqslant \int_0^1 \rho(f - \sigma_k)^2. \tag{2}$$

The last integral can be written as

$$\int_0^1 \rho(f - \bar{f} + \bar{f} - \sigma_k)^2,$$

so that

$$\begin{aligned} 0 &\leqslant \int_0^1 \rho(f - s_k)^2 \\ &\leqslant \int_0^1 \rho(f - \bar{f} + \bar{f} - \sigma_k)^2 \\ &\leqslant \left\{ \left[\int_0^1 \rho(f - \bar{f})^2 \right]^{\frac{1}{2}} + \left[\int_0^1 \rho(\bar{f} - \sigma_k)^2 \right]^{\frac{1}{2}} \right\}^2 \text{ by the Triangle Inequality.} \end{aligned}$$

Therefore, taking the limit as $k \rightsquigarrow \infty$,

$$0 \leqslant \lim_{k \rightsquigarrow \infty} \int_0^1 \rho(f - s_k)^2 \leqslant \int_0^1 \rho(f - \bar{f})^2 < \varepsilon,$$

by (1), where we have used the fact that

$$\lim_{k \rightsquigarrow \infty} \int_0^1 \rho(\bar{f} - \sigma_k)^2 = 0,$$

by *W: page 166, lines 3 and 4*. Since ε is arbitrary, we must have

$$\lim_{k \rightsquigarrow \infty} \int_0^1 \rho(f - s_k)^2 = 0,$$

as required.

THEOREM B

If f is continuous on $[0, 1]$, $f(0) = f(1) = 0$ and

$$\int_0^1 (pf'^2 + qf^2)$$

is finite, then the Fourier series

$$\sum_1^\infty c_n u_n$$

converges uniformly to f.

Proof

Define the partial sums

$$s_k(x) = \sum_1^k c_n u_n(x)$$

and

$$\sigma_k = \sum_1^k \lambda_n c_n^2 \int_0^1 \rho u_n^2. \tag{3}$$

Then Equation (36.15) becomes, after taking the square root of both sides, for all $M > 0$ and for all $N > M$,

$$|s_N(x) - s_M(x)| \leqslant (|G(x, x)|\,|\sigma_N - \sigma_M|)^{\frac{1}{2}}. \tag{4}$$

Now $G(x, y)$ is bounded; in particular, there exists a number A such that

$$|G(x, x)| \leqslant A \qquad 0 \leqslant x \leqslant 1. \tag{5}$$

Also the series (3) converges as $k \rightsquigarrow \infty$ (*W: page 166*), and so given any number $\varepsilon > 0$, there exists N_0 independent of x such that

$$|\sigma_N - \sigma_M| < \frac{\varepsilon^2}{A} \quad \text{whenever } N > M > N_0. \tag{6}$$

Equations (5) and (6) together with (4) give

$$|s_N(x) - s_M(x)| < \left(A\frac{\varepsilon^2}{A}\right)^{\frac{1}{2}} = \varepsilon$$

whenever

$$0 \leqslant x \leqslant 1 \quad \text{and} \quad N > M > N_0.$$

Therefore $\{s_k(x)\}$ is a Cauchy sequence for every x and so converges on $[0, 1]$ to $S(x)$, say. The convergence is uniform, by Cauchy's Test (*Unit 6*) since N_0 is independent of x on $[0, 1]$. Therefore S is continuous, by the properties of uniform convergence, since each s_k is continuous. There remains the question of whether the function S which is the sum of the series is in fact equal to f. We know (Theorem A) that the series $\{s_k\}$ converges in the mean to f, and we shall prove the following:

given that f is continuous,

$$\lim_{k \rightsquigarrow \infty} \int_0^1 \rho(f - s_k)^2 = 0 \tag{7}$$

and

$$\lim_{k \rightsquigarrow \infty} s_k(x) = S(x) \text{ uniformly on } [0, 1], \tag{8}$$

then

$$S(x) = f(x) \qquad x \in [0, 1].$$

To prove this we consider the integral

$$\int_0^1 \rho(f - S)^2 = \int_0^1 \rho[(f - s_k) + (s_k - S)]^2$$

where S is continuous and therefore integrable. Hence, by the Triangle Inequality,

$$0 \leqslant \int_0^1 \rho(f - S)^2 \leqslant \left\{\left[\int_0^1 \rho(f - s_k)^2\right]^{\frac{1}{2}} + \left[\int_0^1 \rho(s_k - S)^2\right]^{\frac{1}{2}}\right\}^2.$$

By conditions (7) and (8) the right-hand side tends to zero as $k \rightsquigarrow \infty$. Therefore

$$\int_0^1 \rho(f - S)^2 = 0,$$

and so

$$f(x) = S(x) \qquad 0 \leqslant x \leqslant 1.$$

For suppose that $f(x) \neq S(x)$ at some point x. Then by the continuity of f and S, $\rho(f - S)^2 > 0$ in some interval containing x, and $\rho(f - S)^2 \geqslant 0$ elsewhere, so that

$$\int_0^1 \rho(f - S)^2 > 0,$$

contradicting the equality above. Therefore $f(x) = S(x)$ at every point on $[0, 1]$.

COROLLARY

$\sum_1^\infty c_n u_n(x)$ is absolutely convergent, and moreover $\sum_1^\infty |c_n|\,|u_n(x)|$ is uniformly convergent.

Proof

The absolute convergence of the series

$$\sum_1^\infty c_n u_n(x)$$

can be proved on the same lines. Simply notice that (36.15) is still valid if

$$\left(\sum_{M+1}^{N} |c_n|\,|u_n(x)|\right)^2$$

is substituted for the left-hand side. It follows that

$$\sum_1^\infty |c_n|\,|u_n(x)|$$

is *uniformly convergent*. (This result is not true unless $f(0) = f(1) = 0$.)

These results are needed to prove the following result (*W: page 168, line 5*).

THEOREM C

Let $\int_0^1 |F|$ be finite Then the problem

$$(pw')' - qw = -F,$$

with

$$w(0) = w(1) = 0$$

has the solution

$$w(x) = \sum_1^\infty \frac{c_n}{\lambda_n} u_n(x),$$

where u_n and λ_n are the eigenfunctions and eigenvalues of the problem

$$(pu')' - qu + \lambda\rho u = 0,$$

$$u(0) = u(1) = 0,$$

for some continuous positive function ρ on [0, 1], and the coefficients $\{c_n\}$ are defined by

$$\frac{F(x)}{\rho(x)} \sim \sum_1^\infty c_n u_n(x).$$

The series for w converges absolutely and uniformly. (Note that many expansions of the solution are possible depending on the choice of ρ.)

Proof

Let G be the Green's function for the system; then the solution of our problem is given by

$$w(x) = \int_0^1 G(x, \xi) F(\xi)\, d\xi.$$

Using the bilinear formula for G we have

$$w(x) = \int_0^1 F(\xi) \left(\sum_1^\infty \frac{u_n(x) u_n(\xi)}{\lambda_n \int_0^1 \rho u_n^2} \right) d\xi \tag{9}$$

for each $x \in [0, 1]$. By Theorem B, the given expansion converges to $G(x, \xi)$ uniformly on $0 \leqslant \xi \leqslant 1$, since the function $\xi \longmapsto G(x, \xi)$ is continuous, piecewise smooth (so that it satisfies the integral condition in Theorem B) and zero at the end points (*W: page 122*). It is known that in this case we can interchange the integral and the summation in (9), giving

$$\begin{aligned} w(x) &= \sum_1^\infty \left[\int_0^1 F(\xi) \frac{u_n(x) u_n(\xi)}{\lambda_n \int_0^1 \rho u_n^2} \right] d\xi \\ &= \sum_1^\infty \frac{u_n(x)}{\lambda_n \int_0^1 \rho u_n^2} \int_0^1 \rho(\xi) \left[\frac{F(\xi)}{\rho(\xi)} \right] u_n(\xi)\, d\xi \\ &= \sum_1^\infty \frac{c_n}{\lambda_n} u_n(x), \end{aligned}$$

according to the definition of c_n. It is clear that this expression is the Fourier series for w. Now, w is continuous and vanishes at the points 0, 1 and

$$\begin{aligned} \int_0^1 (pw'^2 + qw^2) &= -\int_0^1 w[(pw')' - qw] \qquad \text{integrating by parts} \\ &= \int_0^1 wF, \end{aligned}$$

which is finite since w is continuous on the closed interval [0, 1] and

$$\int_0^1 |F|$$

is finite. Hence, by Theorem B and its corollary, the series for w converges absolutely and uniformly.

13.2.3 Vibration of a Variable String

It is rather hard to find equations of the form (36.1) which have non-constant coefficients but which can be solved in terms of elementary functions to display the eigenfunctions and eigenvalues we have been talking about. We shall now meet an example which, though artificial, is one that can be worked out in full. In *Unit 14, Bessel Functions* another example is given which requires us to explore completely new functions.

READ W: Section 37, pages 169 to 171.

Notes

(i) *W: page 169, Equation (37.2)*
Here $p \equiv 1$, $q \equiv 0$, $\rho(x) = (1 + x)^{-2}$.

(ii) *W: page 169, lines −12 and −11*
We test whether there is a value of a satisfying the equation by putting

$$X(x) = (1 + x)^a$$

into (37.2). Then a must satisfy

$$a(a - 1)(1 + x)^{a-2} + \lambda(1 + x)^{a-2} = 0,$$

for all x. This is equivalent to

$$a(a - 1) + \lambda = 0.$$

(iii) *W: page 170, line 5*

We have

$$(1 + x)^{\frac{1}{2}(1 + i\sqrt{4\lambda - 1})} = (1 + x)^{\frac{1}{2}}(1 + x)^{\frac{1}{2}i\sqrt{4\lambda - 1}},$$

where $\frac{1}{2}\sqrt{4\lambda - 1}$ is real. Now, a^{ib}, when a and b are real, is *defined* in terms of the complex exponential function as $e^{ib \log a}$. Here $a = 1 + x$, $b = \frac{1}{2}\sqrt{4\lambda - 1}$.

(iv) *W: page 170, line 8*
Formally, we write the complex solution given in *lines 6* and *7* in the form $X = X_1 + iX_2$, where

$$X_1(x) = (1 + x)^{\frac{1}{2}} \cos\left(\sqrt{(\lambda - \tfrac{1}{4})} \log(1 + x)\right),$$

$$X_2(x) = (1 + x)^{\frac{1}{2}} \sin\left(\sqrt{(\lambda - \tfrac{1}{4})} \log(1 + x)\right),$$

X_1 and X_2 being real functions. Then, since X is a solution,

$$0 = X''(x) + \frac{\lambda}{(1 + x)^2} X(x)$$

$$= \left[X_1''(x) + \frac{\lambda}{(1 + x)^2} X_1(x)\right] + i\left[X_2''(x) + \frac{\lambda}{(1 + x)^2} X_2(x)\right].$$

Each of the brackets on the right is real, and therefore the complex number on the right is zero if and only if the brackets are separately zero. Therefore X_1 and X_2 are, separately, solutions of the original differential equation. Strictly speaking, the correct thing to do at this stage would be to verify that the real functions X_1 and X_2 are two independent solutions of the differential equation (37.2).

13.3 FURTHER PROPERTIES OF EIGENVALUES AND EIGENFUNCTIONS

13.3.0 Introduction

In this section we shall meet a number of important theorems regarding the eigenvalues of boundary value problems and the zeros of their eigenfunctions. The two results concerning the eigenfunctions are the *Separation Theorem* and the *Oscillation Theorem*, and the result about the eigenvalues is the *Monotonicity Theorem.*

We begin the discussion by proving a comparison result which will turn out eventually to be a particular case of the Monotonicity Theorem. We then discover that the first eigenfunction of a boundary value problem has no zeros between the end points, i.e. it does not change sign on the interval. Finally we plough through the main theorems.

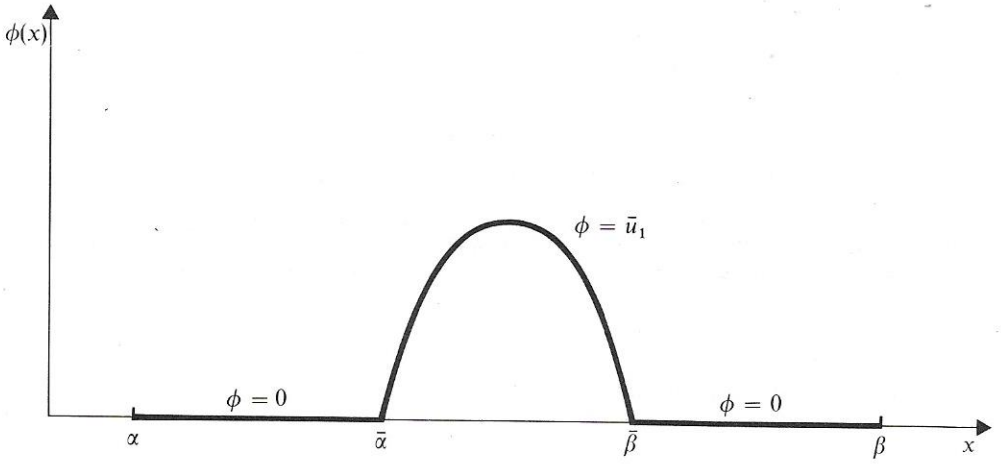
13.3.1 A Comparison Theorem for Sturm–Liouville Problems

READ W: page 172, line 13 to page 173, line 7.

We have not asked you to read the first two paragraphs of *W: Section 38* because the argument is difficult to follow. We shall however obtain Weinberger's result in Section 13.3.2.

Notes

(i) *W: page 172, line −4*
Note the shape of the trial function, ϕ.



(ii) *W: page 173, lines 3 to 5*
The ratio referred to is

$$\frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2)}{\int_\alpha^\beta \rho\phi^2}.$$

The reasoning is not easy to follow, but is credible. The strict inequality

$$\bar{\lambda}_1 > \lambda_1$$

is important.

General Comment

The two arguments which we have described as difficult to follow (*W: page 172, line 4* and *W: page 173, line 3*) are valid, but they rely on a slight extension of the result in *W: page 163*. Effectively, it can be shown that if ϕ is continuous and ϕ'' is piecewise continous on $[\alpha, \beta]$,

$$\phi(\alpha) = \phi(x_1) = \cdots = \phi(x_m) = \phi(\beta) = 0$$

where $x_1, \ldots, x_m$ are the points of discontinuity of ϕ'' and the Rayleigh quotient $R[\phi] = \lambda_1$ then ϕ is an eigenfunction corresponding to λ_1.

13.3.2 The Zeros of Eigenfunctions

In *W: page 172* (the first paragraph), it is shown that the first eigenfunction, u_1, does not vanish on (α, β). To express this another way,

the first eigenfunction of the system

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (\alpha, \beta),$$
$$u(\alpha) = u(\beta) = 0,$$

is of one sign throughout the interval.

Of course, we can change this sign by multiplying the function by a constant.

We shall see in the proof of the Oscillation Theorem that the kth eigenfunction has at most $k - 1$ zeros; thus our result is just a special case of this.

It follows that

any other eigenfunction changes sign on (α, β),

for if $k > 1$, $\int_\alpha^\beta \rho u_k u_1 = 0$; or alternatively

if α *and* β *are consecutive zeros of any solution of the differential equation*

$$(pu')' - qu + \lambda\rho u = 0$$

with any given λ, λ *must be the lowest eigenvalue for the interval* $[\alpha, \beta]$.

READ W: page 173, line 8 to page 175, line −11 (the end of the section).

Notes

(i) *W: page 173, SEPARATION THEOREM*
The proof given is not complete, since it assumes that u has at least two zeros. Substitute the following proof.

Proof

The proof is by contradiction. Suppose that u does not vanish on the open interval $(\bar{\alpha}, \bar{\beta})$; then u and v each has constant sign on $(\bar{\alpha}, \bar{\beta})$ and we may assume without loss of generality that $u > 0$, $v > 0$ on $(\bar{\alpha}, \bar{\beta})$. Multiply Equation (38.4) by v and Equation (38.5) by u and subtract, then integrate from $\bar{\alpha}$ to $\bar{\beta}$ (compare Lagrange's Identity, *Unit 9*). We obtain

$$\int_{\bar{\alpha}}^{\bar{\beta}} [v(pu')' - u(pv')'] = (\bar{\lambda} - \lambda) \int_{\bar{\alpha}}^{\bar{\beta}} \rho uv \leqslant 0, \tag{1}$$

with strict inequality if $\bar{\lambda} < \lambda$.

Now integrate the left-hand side by parts. We obtain

$$\Big[v(pu') - u(pv')\Big]_{\bar{\alpha}}^{\bar{\beta}} \leqslant 0,$$

and since $v(\bar{\alpha}) = v(\bar{\beta}) = 0$ by hypothesis, this becomes

$$p(\bar{\alpha})u(\bar{\alpha})v'(\bar{\alpha}) - p(\bar{\beta})u(\bar{\beta})v'(\bar{\beta}) \leqslant 0. \tag{2}$$

Since $v > 0$ on $(\bar{\alpha}, \bar{\beta})$ and $v(\bar{\alpha}) = v(\bar{\beta}) = 0$, we have $v'(\bar{\alpha}) > 0$ and $v'(\bar{\beta}) < 0$. Now suppose firstly that $u(\bar{\alpha})$ and $u(\bar{\beta})$ are not both zero. Since $p(x) > 0$ on $[\bar{\alpha}, \bar{\beta}]$ (*W: page 160, line 12*), and $u(x) \geqslant 0$ on $[\bar{\alpha}, \bar{\beta}]$, the left-hand side of (2) is positive. (2) is therefore contradicted, so u must change sign, and therefore vanish, somewhere on $(\bar{\alpha}, \bar{\beta})$.

Secondly, suppose that $u(\bar{\alpha}) = u(\bar{\beta}) = 0$ and $\lambda \neq \bar{\lambda}$ (in this case u and v are independent eigenfunctions for $[\bar{\alpha}, \bar{\beta}]$). Then (2) becomes

$$p(\bar{\alpha})u(\bar{\alpha})v'(\bar{\alpha}) - p(\bar{\beta})u(\bar{\beta})v'(\bar{\beta}) < 0, \tag{3}$$

(see Equation (1)), and again there is a contradiction, since the left-hand side is zero.

Finally suppose that $u(\bar{\alpha}) = u(\bar{\beta}) = 0$, and $\lambda = \bar{\lambda}$. Then u and v are eigenfunctions for the interval $[\bar{\alpha}, \bar{\beta}]$, with the same eigenvalue, and are therefore multiples of each other (see SAQ 3).

(ii) *W: page 174, lines 1 to 9*
We expand the argument in this note.

Firstly, let $\{c_m\}$ be *any* set of constants, not necessarily satisfying the orthogonality conditions on *W: page 173, lines −4 and −3*. Then,

$$\int_\alpha^\beta [p\phi'^2 + q\phi^2] = \int_\alpha^\beta \left[p \left(\sum_{m=1}^{l} c_m u^{(m)\prime} \right)^2 + q \left(\sum_{m=1}^{l} c_m u^{(m)} \right)^2 \right]$$

$$= \int_\alpha^\beta \left[p \sum_{m=1}^{l} \sum_{n=1}^{l} c_m c_n u^{(m)\prime} u^{(n)\prime} + q \sum_{m=1}^{l} \sum_{n=1}^{l} c_m c_n u^{(m)} u^{(n)} \right].$$

Since $u^{(m)}u^{(n)} = u^{(m)\prime}u^{(n)\prime} = 0$ for $m \neq n$, this is equal to

$$\int_\alpha^\beta \left[p \sum_{m=1}^{l} c_m^2 (u^{(m)\prime})^2 + q \sum_{m=1}^{l} c_m^2 (u^{(m)})^2 \right],$$

and since $u^{(m)}(x) = 0$ except on (x_{m-1}, x_m) where it equals $u_k(x)$, this is in turn equal to

$$\sum_{m=1}^{l} c_m^2 \int_{x_{m-1}}^{x_m} [pu_k'^2 + qu_k^2].$$

The argument continues up to *line 4*. Therefore, finally,

$$\lambda_k = \frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2)}{\int_\alpha^\beta \rho\phi^2}. \tag{1}$$

Now, (1) holds for *any* set of constants c_m $(m = 1, 2, \ldots, l)$. We can therefore choose these l constants to satisfy the $l - 1$ linear equations *W: page 173, lines −4 and −3*,

$$\int_\alpha^\beta \rho\phi u_1 = 0, \ldots, \int_\alpha^\beta \rho\phi u_{l-1} = 0; \tag{2}$$

that is, we can arrange for ϕ to be orthogonal to the first $l - 1$ eigenfunctions, without disturbing the truth of (1). But the absolute minimum of

$$\frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2)}{\int_\alpha^\beta \rho\phi^2}$$

over *all* admissible ϕ subject to (2) is, by the minimum principle, equal to λ_l. Therefore

$$\lambda_l \leqslant \lambda_k,$$

and so $l \leqslant k$. Now the assumed number of zeros of u_k on the open interval (α, β) is equal to $l - 1$ (*W: page 173, line −11*). Therefore the number of zeros of u_k on (α, β) is not more than $k - 1$. But we know that the number of zeros is at least $k - 1$, so that equality must hold (*W: page 173, line −14*).

Note that, in particular, it follows that the first eigenfunction has no zeros in (α, β).

SAQ 15

Let u_k, u_{k+1} be successive eigenfunctions of the system

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (0, 1)$$

$$u(0) = u(1) = 0.$$

Show that there is exactly one zero of u_k on the open interval between every two zeros of u_{k+1} in (0, 1), and exactly one zero of u_{k+1} between every two zeros of u_k (i.e. the zeros interlace).

(Solution on p. 36.)

SAQ 16

Let u^*, u^{**} be any two linearly independent solutions of the equation

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (0, 1).$$

Prove that on the open interval between any two successive zeros of u^* there lies exactly one zero of u^{**}, and on the open interval between any two successive zeros of u^{**} there lies exactly one zero of u^*.

(Solution on p. 36.)

SAQ 17

Confirm that the nth eigenfunction, X_n, of the variable string problem in *W: Section 37* has exactly $n - 1$ zeros in (0, 1). Check that the zeros of X_n and X_{n+1} interlace (see SAQ 15).

(Solution on p. 36.)

SAQ 18

W: page 175; Exercise 1.

(Solution on p. 36.)

SAQ 19 (*Difficult*)

(a) Prove that, if $q \equiv 0$, reducing the interval (α, β), increasing p, q or b, or decreasing ρ increases the lowest eigenvalue of the system (38.6). (You may assume that the minimum principle (38.7) applies.)

(b) *W: page 175, Exercise 2.*

(Solution on p. 37.)

SAQ 20 (*Difficult*)

Prove that the kth eigenfunction for the system (38.6) has at most $k - 1$ zeros on the open interval $\alpha < x < \beta$.

(Solution on p. 38.)

SAQ 21 (*Difficult*)

Prove in full the result in *W: page 175, lines 16 and 17*, that the eigenvalues of system (38.6) separate those of (38.1).

(Solution on p. 39.)

SAQ 22

W: page 176, Exercise 3.
(The meaning of the result is that λ_k behaves like $k^2\pi^2$ for large k.)

HINT: Compare the system with a system having constant coefficients.

(Solution on p. 40.)

SAQ 23

Consider the system

$$u'' + \lambda\rho u = 0 \qquad \text{in } (\alpha, \beta),$$

$$u(\alpha) = u(\beta) = 0.$$

(a) Prove that

$$\frac{k^2\pi^2}{(\beta-\alpha)^2\rho_{\max}} \leqslant \lambda_k \leqslant \frac{k^2\pi^2}{(\beta-\alpha)^2\rho_{\min}}$$

where $\rho_{\max}$, $\rho_{\min}$ represent the maximum and minimum values of ρ respectively on $[\alpha, \beta]$.

(b) Let the kth eigenfunction for the system have zeros at the points $\alpha = x_0, x_1, x_2, \ldots, x_k = \beta$. Prove that for $i = 0, 1, \ldots, k-1$,

$$\frac{\pi^2}{(x_{i+1}-x_i)^2\rho_{i,\max}} \leqslant \lambda_k \leqslant \frac{\pi^2}{(x_{i+1}-x_i)^2\rho_{i,\min}},$$

where $\rho_{i,\max}$ and $\rho_{i,\min}$ are the maximum and minimum values respectively of $\rho(x)$ on the interval $[x_i, x_{i+1}]$. Deduce that

$$\rho_{i,\min}^{\frac{1}{2}}(x_{i+1}-x_i) \leqslant \frac{\pi}{\lambda_k^{\frac{1}{2}}} \leqslant \rho_{i,\max}^{\frac{1}{2}}(x_{i+1}-x_i),$$

for $i = 0, 1, \ldots, k-1$.

(c) Let M_k be the maximum distance between successive zeros of the kth eigenfunction; that is, the maximum value of $(x_{i+1}-x_i)$, $i = 0, 1, \ldots, k-1$. Assuming that $\lim_{k\to\infty} M_k = 0$, prove that

$$\lim_{k\to\infty}\frac{\lambda_k}{k^2} = \frac{\pi^2}{\left[\int_\alpha^\beta \rho^{\frac{1}{2}}(x)\,dx\right]^2}.$$

(Solution on p. 40.)

13.4 SUMMARY

This unit is concerned mainly with the properties of the eigenvalues and eigenfunctions of the self-adjoint system

$$(pu')' - qu + \lambda\rho u = 0 \qquad \text{in } (\alpha, \beta), \tag{1}$$
$$u(\alpha) = u(\beta) = 0,$$

and with related systems having boundary conditions involving $u'(\alpha)$ and $u'(\beta)$; p, q and ρ are continuous functions on $[a, b]$, p has a continuous first derivative on (α, β), and p and ρ are positive and q nonnegative on $[\alpha, \beta]$.

Some of the results in W are stated for the standard interval $[0, 1]$. Since they apply without change to an arbitrary interval $[\alpha, \beta]$, we quote them for this general interval. The eigenvalues $\lambda_1, \lambda_2, \ldots$, are assumed to be indexed in order of size.

THE MINIMUM PRINCIPLE

Let ϕ be a continuous, piecewise smooth function on $[\alpha, \beta]$, and let $\phi(\alpha) = \phi(\beta) = 0$. Then

$$\lambda_1 \leqslant \frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2)}{\int_\alpha^\beta \rho\phi^2}, \tag{2}$$

where λ_1 is the lowest eigenvalue of the system (1). If ϕ is allowed to range over all continuous piecewise smooth functions, then the minimum of the right-hand side is attained when $\phi = u_1$, where u_1 is an eigenfunction corresponding to λ_1, and then equality holds in Equation (2).

For the other eigenvalues we have the following extension:

$$\lambda_k = \min \frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2)}{\int_\alpha^\beta \rho\phi^2},$$

the minimum being over functions ϕ which satisfy the conditions

(a) ϕ is continuous and piecewise smooth on $[\alpha, \beta]$,

(b) $\phi(\alpha) = \phi(\beta) = 0$,

(c) $\int_\alpha^\beta \rho\phi u_1 = \int_\alpha^\beta \rho\phi u_2 = \cdots = \int_\alpha^\beta \rho\phi u_{k-1} = 0$, where $u_1, \ldots, u_k$ are eigenfunctions corresponding to the first $k - 1$ eigenvalues.

Any minimizing function of this class which is twice continuously differentiable on (α, β) is an eigenfunction.

Properties of the eigenvalues and eigenfunctions of the system (1)

The eigenvalues are positive, infinite in number and discrete (this justifies indexing them in order).

Eigenfunctions belonging to a single eigenvalue are linearly dependent.

Eigenfunctions corresponding to different eigenvalues are orthogonal with weight function ρ, i.e.

$$\int_\alpha^\beta \rho u_k u_l = 0, \qquad k \neq l.$$

MONOTONICITY THEOREM

Reducing the interval $[\alpha, \beta]$ to a subinterval of $[\alpha, \beta]$, increasing p or q, or decreasing ρ increases all the eigenvalues.

SEPARATION THEOREM

Let u and v be *any* two solutions of the equations

$$(pu')' - qu + \lambda\rho u = 0,$$

$$(pv')' - qv + \bar{\lambda}\rho v = 0$$

respectively, with $\bar{\lambda} \leqslant \lambda$. Then if $\bar{\alpha}$ and $\bar{\beta}$ are consecutive zeros of v, there is at least one zero of u in the open interval $(\bar{\alpha}, \bar{\beta})$, unless $\bar{\lambda} = \lambda$ and v is a multiple of u.

OSCILLATION THEOREM

The kth eigenfunction u_k has exactly $k - 1$ zeros in the open interval (α, β). The zeros of successive eigenfunctions interlace.

THEOREM A

The set of eigenfunctions $\{u_n\}$ is complete for the space of functions f for which $\int_\alpha^\beta \rho f^2$ is finite; i.e. for such functions

$$\lim_{k\to\infty} \int_\alpha^\beta \rho\left(f - \sum_1^k c_n u_n\right)^2 = 0$$

where $\{c_n\}$ is the sequence of Fourier coefficients of f with respect to $\{u_n\}$.

THEOREM B

If f is a continuous function such that $f(\alpha) = f(\beta) = 0$, and

$$\int_\alpha^\beta (pf'^2 + qf^2)$$

exists, its Fourier series in terms of the complete set of eigenfunctions $\{u_n\}$ converges uniformly to f on $[\alpha, \beta]$.

We also considered briefly the more general problem

$$\begin{aligned} &(pv')' - qv + \mu\rho v = 0 \qquad \text{in } (\alpha, \beta), \\ &v(\alpha) = 0, \\ &p(\beta)v'(\beta) + bv(\beta) = 0, \end{aligned} \tag{3}$$

whose eigenvalues are given by the minimum principle

$$\mu_k = \min \frac{\int_\alpha^\beta (p\phi'^2 + q\phi^2) + b\phi^2(\beta)}{\int_\alpha^\beta \rho\phi^2}$$

the minimum being taken over all functions ϕ which satisfy the conditions

(a) ϕ is continuous and piecewise smooth on $[\alpha, \beta]$,

(b) $\phi(\alpha) = p(\beta)\phi'(\beta) + b\phi(\beta) = 0$,

(c) $\int_\alpha^\beta \rho\phi v_1 = \int_\alpha^\beta \rho\phi v_2 = \cdots = \int_\alpha^\beta \rho\phi v_{k-1} = 0$, where $v_1, \ldots, v_{k-1}$ are the eigenfunctions corresponding to the first $k - 1$ eigenvalues.

Any minimizing function of this class which is twice continuously differentiable on (α, β) is an eigenfunction v_k of the system (3).

The kth eigenfunction has exactly $k - 1$ zeros in (α, β).

Increasing p or q or b, or decreasing ρ, or taking a subinterval of $[\alpha, \beta]$ increases the eigenvalues of (3).

The eigenvalues of the system (3) separate those of the system (1) and vice versa.

13.5 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

The general solution of the equation is of the form

$$u(x) = A\cos\lambda^{\frac{1}{2}}x + B\sin\lambda^{\frac{1}{2}}x,$$

and the boundary conditions give

$$A\cos\lambda^{\frac{1}{2}}\alpha + B\sin\lambda^{\frac{1}{2}}\alpha = 0 \tag{1}$$

and

$$A\cos\lambda^{\frac{1}{2}}\beta + B\sin\lambda^{\frac{1}{2}}\beta = 0. \tag{2}$$

These equations have a nonzero solution provided that the determinant of the coefficients is zero, i.e. if

$$\cos\lambda^{\frac{1}{2}}\alpha\sin\lambda^{\frac{1}{2}}\beta - \sin\lambda^{\frac{1}{2}}\alpha\cos\lambda^{\frac{1}{2}}\beta = 0$$

or

$$\sin\lambda^{\frac{1}{2}}(\beta - \alpha) = 0. \tag{3}$$

Therefore the eigenvalues are given by

$$\lambda = \lambda_n = \frac{n^2\pi^2}{(\beta-\alpha)^2} \qquad n = 1, 2, \ldots . \tag{4}$$

The solution $\lambda = 0$ of (3) is rejected since it gives $A = 0$ and therefore the trivial solution $u = 0$. The solutions corresponding to $n = -1, -2, \ldots$ merely duplicate (4), and so are omitted. To find the corresponding eigenfunctions we solve (1) and (2) with $\lambda = \lambda_n$; there will clearly be an infinite number of solutions, for which we have, in general,

$$\frac{A}{B} = -\tan\lambda_n^{\frac{1}{2}}\alpha = -\tan\frac{n\pi\alpha}{\beta-\alpha},$$

from Equation (1), which is the same as $-\tan\lambda_n^{\frac{1}{2}}\beta$ obtained from Equation (2). The eigenfunctions are therefore given by

$$\begin{aligned} u_n(x) &= C\left[-\sin\frac{n\pi\alpha}{\beta-\alpha}\cos\frac{n\pi x}{\beta-\alpha} + \cos\frac{n\pi\alpha}{\beta-\alpha}\sin\frac{n\pi x}{\beta-\alpha}\right] \\ &= C\sin\frac{n\pi(x-\alpha)}{\beta-\alpha} \end{aligned} \tag{5}$$

where C is an arbitrary constant. The end point α has no special standing in (5); an alternative process would have given

$$u_n(x) = D\sin\frac{n\pi(x-\beta)}{\beta-\alpha},$$

which is the same as (5).

Solution to SAQ 2

The general solution is of the form

$$A\cos\lambda^{\frac{1}{2}}x + B\sin\lambda^{\frac{1}{2}}x,$$

and the boundary conditions give

$$-A\lambda^{\frac{1}{2}}\sin\lambda^{\frac{1}{2}}\alpha + B\lambda^{\frac{1}{2}}\cos\lambda^{\frac{1}{2}}\alpha = 0 \tag{1}$$

and

$$-A\lambda^{\frac{1}{2}}\sin\lambda^{\frac{1}{2}}\beta + B\lambda^{\frac{1}{2}}\cos\lambda^{\frac{1}{2}}\beta = 0. \tag{2}$$

If A and B are not both zero, we must have

$$\lambda(-\sin \lambda^{\frac{1}{2}}\alpha \cos \lambda^{\frac{1}{2}}\beta + \sin \lambda^{\frac{1}{2}}\beta \cos \lambda^{\frac{1}{2}}\alpha) = 0,$$

or

$$\lambda \sin \lambda^{\frac{1}{2}}(\beta - \alpha) = 0.$$

The eigenvalues are given by

$$\lambda = \lambda_n = \frac{n^2\pi^2}{(\beta - \alpha)^2} \qquad n = 0, 1, 2, \ldots .$$

We do not reject $\lambda = 0$, since this does not yield the trivial solution in this case. In fact, corresponding to the eigenvalue $\lambda_0 = 0$, we have the solution

$$u_0 = A, \tag{3}$$

where A is any constant.

For the other eigenfunctions we put $\lambda = \lambda_n$ ($n \geqslant 1$) and solve (1) and (2), obtaining

$$u_n(x) = C \cos \frac{n\pi(x - \alpha)}{\beta - \alpha}$$

where C is arbitrary (or $D \cos [n\pi(x - \beta)/(\beta - \alpha)]$ where D is arbitrary).

To resolve the paradox of $\lambda = 0$ being an eigenvalue, we follow through the argument of *W: page 161* for our problem. Equation (36.4) is still true (λ being any eigenvalue and u a corresponding eigenfunction) with a suitable change in the limits, because

$$puu' \Big]_{\alpha}^{\beta} = 0.$$

The function q *is* nonnegative for our case—in fact $q = 0$. The difference lies in the fact that

$$u'(x) = 0 \qquad \alpha < x < \beta,$$

is now a possibility; indeed u given by Equation (3) satisfies this condition. Therefore the numerator of Equation (36.4) can be zero and the eigenvalue $\lambda = 0$ is correctly given.

Solution to SAQ 3

Suppose that there exist two linearly independent eigenfunctions corresponding to λ_k, $u_k^{(1)}$ and $u_k^{(2)}$. Let $u = U$ be a particular solution of the equation

$$(pu')' - qu + \lambda_k \rho u = 0, \tag{1}$$

for which $U(\alpha) = 1$, but which is otherwise unrestricted. The existence of such a solution is guaranteed by the Existence Theorem (*Unit M201 33, Existence and Uniqueness Theorem*). Then, since $u_k^{(1)}$ and $u_k^{(2)}$ are by hypothesis linearly independent solutions of the second order differential equation, they form a basis for the solution space and there exist constants A and B such that

$$U(x) = Au_k^{(1)}(x) + Bu_k^{(2)}(x) \qquad x \in [\alpha, \beta].$$

But the right-hand side is zero when $x = \alpha$, and the left-hand side has the value 1. Therefore we have a contradiction, and we conclude that any two eigenfunctions corresponding to the same eigenvalue are linearly dependent.

Solution to SAQ 4

Let λ, μ be two different eigenvalues, and u, v a pair of corresponding eigenfunctions. We carry out the procedure in *W: pages 160 to 161*; u and v satisfy

$$(pu')' - qu + \lambda\rho u = 0,$$

$$(pv')' - qv + \mu\rho v = 0.$$

We multiply the first equation by v and the second by u, subtract, and integrate over $[\alpha, \beta]$; after cancellation we have

$$\int_\alpha^\beta [v(pu')' - u(pv')'] + (\lambda - \mu)\int_\alpha^\beta \rho uv = 0.$$

After integrating by parts (an alternative view to that of *W: page 161, line 4*) and rearranging, we have

$$(\lambda - \mu)\int_\alpha^\beta \rho uv = \Big[p(vu' - uv')\Big]_\alpha^\beta. \tag{1}$$

We aim to show that the bracket is zero. Since $A \neq 0$ we can write

$$p(vu' - uv') = \frac{1}{A}p[v(Au' + Bu) - u(Av' + Bv)],$$

which is zero at $x = \alpha$, since u and v satisfy the boundary conditions. Similarly we may show that it is zero at $x = \beta$. Therefore (1) gives

$$(\lambda - \mu)\int_\alpha^\beta \rho uv = 0,$$

and since $\lambda \neq \mu$,

$$\int_\alpha^\beta \rho uv = 0.$$

Solution to SAQ 5

Such periodic boundary conditions arise when, for example, θ represents the angular coordinate in plane polar coordinates. When $\theta = \pi$ we start "going round again", and the boundary conditions state that the function and its first derived function (and, because of the differential equation, all its derived functions) are continuous across $\theta = \pi$. The particular equation in our problem arises from separating Laplace's equation in polar coordinates (*W: Section 24*).

The general solution is $u(\theta) = A\cos\lambda^{\frac{1}{2}}\theta + B\sin\lambda^{\frac{1}{2}}\theta$, where A and B are any constants. The boundary conditions give

$$u(-\pi) - u(\pi) = 0 = -2B\sin\lambda^{\frac{1}{2}}\pi \tag{1}$$

and

$$u'(-\pi) - u'(\pi) = 0 = 2A\lambda^{\frac{1}{2}}\sin\lambda^{\frac{1}{2}}\pi. \tag{2}$$

One solution is $\lambda^{\frac{1}{2}} = 0$, which gives the eigenfunction

$$u(\theta) = A,$$

where A is any constant.

If $\lambda \neq 0$, then $\lambda = n^2$ $(n = 1, 2, 3, \ldots)$ satisfies (1) and (2). Both A and B may have any values, and the corresponding eigenfunctions are given by

$$u(\theta) = A\cos n\theta + B\sin n\theta.$$

The simplest basis for the whole space of eigenfunctions corresponding to the eigenvalue $n^2 > 0$ is

$$\{\cos n\theta, \sin n\theta\}.$$

Thus we see that the eigenspace has dimension 2.

Solution to SAQ 6

The general solution is

$$u(x) = A\cos\lambda^{\frac{1}{2}}x + B\sin\lambda^{\frac{1}{2}}x;$$

the boundary conditions give

$$u(0) = 0 = A \tag{1}$$

and

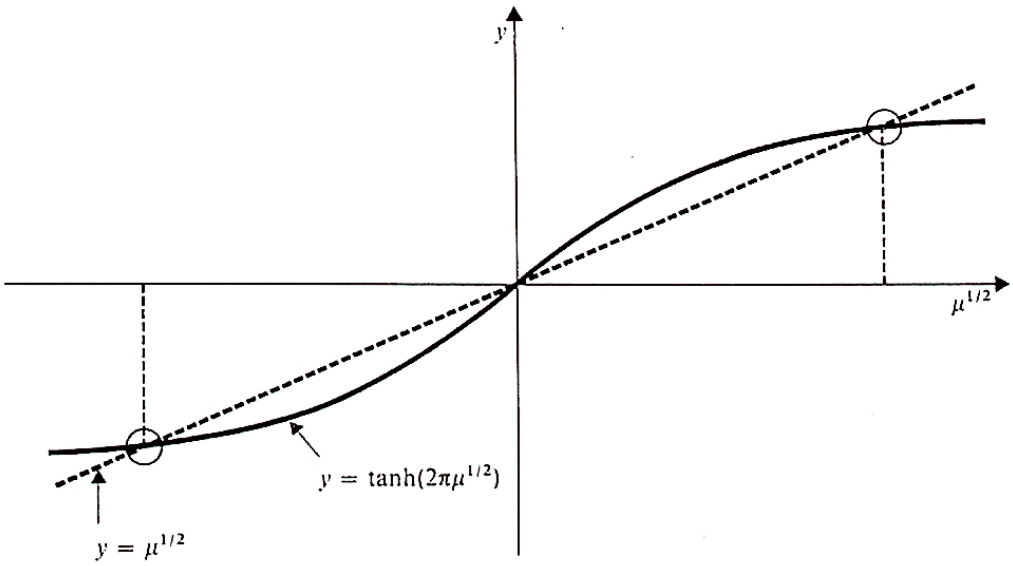
$$u'(2\pi) - u(2\pi) = 0 = B(\lambda^{\frac{1}{2}}\cos\lambda^{\frac{1}{2}}2\pi - \sin\lambda^{\frac{1}{2}}2\pi). \tag{2}$$

Equation (2) gives $\lambda^{\frac{1}{2}} = \tan(2\pi\lambda^{\frac{1}{2}})$, as required. There is an infinite number of positive eigenvalues, as well as $\lambda = 0$, as can be seen by plotting a graph with axes $\lambda^{\frac{1}{2}}$, y, and sketching the line $y = \lambda^{\frac{1}{2}}$ and the curve $y = \tan(2\pi\lambda^{\frac{1}{2}})$.

To find any negative eigenvalues we write $\lambda = -\mu, (\mu > 0)$ and $\lambda^{\frac{1}{2}} = i\mu^{\frac{1}{2}}$; the equation for the roots becomes

$$\mu^{\frac{1}{2}} = \tanh(2\pi\mu^{\frac{1}{2}}),$$

since $\tan ic = i\tanh c$. The sketch shows two values of $\mu^{\frac{1}{2}}$ equal in magnitude and opposite in sign, giving a single negative root λ.



Note that the existence of a negative eigenvalue does not contradict *W: page 161, line −11*, since the boundary conditions in our problem are different and (36.4) does not hold.

Solution to SAQ 7

The first eigenvalue is $\lambda_1 = \pi^2$ and the corresponding eigenfunction $u_1(x)$ is $A\sin\pi x$ (A some constant). The Rayleigh quotient becomes

$$\frac{\int_0^1 A^2\pi^2\cos^2\pi x\,dx}{\int_0^1 A^2\sin^2\pi x\,dx} = \frac{A^2\pi^2\int_0^1 \frac{1}{2}(1+\cos 2\pi x)\,dx}{A^2\int_0^1 \frac{1}{2}(1-\cos 2\pi x)\,dx} = \pi^2,$$

as required.

Solution to SAQ 8

Let ψ be an admissible function which minimizes the given quotient, and let $\tilde{\psi}$ be *any* admissible function. The minimum property implies that

$$R[\tilde{\psi}] \geqslant R[\psi],$$

where R is shorthand for the given Rayleigh-type quotient. We write the comparison function $\tilde{\psi}$ in the form

$$\tilde{\psi} = \psi + \tau\phi,$$

where ϕ is an arbitrary admissible function, and τ is an arbitrary parameter. (Obviously, we do not in any way increase the degree of "arbitrariness" available by introducing τ: we can in fact do without it, but the argument is simpler if we put it in. It is convenient to think of τ being "small" in some sense.) Then, *for any given* ϕ the quotient $R[\psi + \tau\phi]$ has a minimum when $\tau = 0$, that is,

$$\frac{d}{d\tau}\left[\frac{\int_0^1 (\psi' + \tau\phi')^2 - [\psi(1) + \tau\phi(1)]^2}{\int_0^1 (\psi + \tau\phi)^2}\right]_{\tau=0}$$

$$= \frac{\int_0^1 2\psi'\phi' - 2\psi(1)\phi(1)}{\int_0^1 \psi^2} - \frac{\left\{\int_0^1 \psi'^2 - [\psi(1)]^2\right\}\int_0^1 2\psi\phi}{\left(\int_0^1 \psi^2\right)^2} = 0. \quad (1)$$

Let the minimum value of the quotient be equal to μ, i.e.

$$R[\psi] = \frac{\int_0^1 \psi'^2 - [\psi(1)]^2}{\int_0^1 \psi^2} = \mu;$$

then from (1) we obtain

$$\int_0^1 (\psi'\phi' - \mu\psi\phi) - \psi(1)\phi(1) = 0, \quad (2)$$

for every admissible ϕ. We integrate (2) by parts, assuming that ψ is twice continuously differentiable; then

$$0 = -\int_0^1 \phi(\psi'' + \mu\psi) + \psi'(1)\phi(1) - \psi'(0)\phi(0) - \psi(1)\phi(1)$$

$$= -\int_0^1 \phi(\psi'' + \mu\psi) + [\psi'(1) - \psi(1)]\phi(1) - \psi'(0)\phi(0).$$

The terms at the end come to zero, since $\psi'(1) - \psi(1) = 0$ and $\phi(0) = 0$, by the boundary conditions. The rest of the argument is as in *W: page 164*.

Solution to SAQ 9

Suppose that $\lambda_1, \ldots, \lambda_{k-1}$ and $u_1, \ldots, u_{k-1}$ have been found. Then to find λ_k, u_k, we proceed exactly as in *W* with $p \equiv 1$, $q \equiv 0$, $\rho \equiv 1$; integrating over $(-\pi, \pi)$ instead of (0, 1), we obtain, corresponding to Equation (36.9),

$$\int_{-\pi}^{\pi} [\psi'\phi' - \mu\psi\phi] = 0, \quad (1)$$

where ψ is the minimizing function and ϕ an arbitrary admissible function subject to the boundary conditions of the present problem and to the $k - 1$ orthogonality conditions (if $k = 1$ there is no orthogonality condition to be satisfied).

We integrate (1) by parts; for all admissible ϕ we have

$$\int_{-\pi}^{\pi} [-\psi'' - \mu\psi]\phi + \left[\psi'\phi\right]_{-\pi}^{\pi} = 0.$$

Now

$$\left[\psi'\phi\right]_{-\pi}^{\pi} = \psi'(\pi)\phi(\pi) - \psi'(-\pi)\phi(-\pi) = 0,$$

because of the boundary conditions satisfied by ϕ and ψ. The proof continues as in *W: page 164*, only the boundary conditions being different.

We saw in SAQ 5 that the eigenvalues are $\lambda = n^2$ $(n = 0, 1, 2, \ldots)$. If $n = 0$ the eigenspace has dimension 1, and if $n > 0$ it has dimension 2. In the minimization process, the successive minima delivered occur in equal pairs (after the first). The second eigenfunction delivered in any such pair is, of course, different from the first, since it is required to be orthogonal to the first.

Solution to SAQ 10

We proceed exactly as in *W* up to Equation (36.9), obtaining

$$\int_0^1 [p\psi'\phi' + q\psi\phi - \mu\rho\psi\phi] = 0.$$

Now we integrate by parts, so that

$$-\int_0^1 \phi[(p\psi')' - q\psi + \mu\rho\psi] + p(1)\psi'(1)\phi(1) - p(0)\psi'(0)\phi(0) = 0. \qquad (1)$$

and for every admissible ϕ we have $\phi(0) = 0$, so the last term vanishes.

Now in order that (1) should vanish for *every* admissible ϕ, it must vanish in particular for such $\tilde{\phi}$ as satisfy additionally the condition

$$\tilde{\phi}(1) = 0.$$

Therefore

$$\int_0^1 \tilde{\phi}[(p\psi')' - q\psi + \mu\rho\psi] = 0$$

for every $\tilde{\phi}$ for which $\tilde{\phi}(0) = \tilde{\phi}(1) = 0$, and the argument in *W: page 164, lines 5 to 10* shows that ψ must satisfy

$$(p\psi')' - q\psi + \mu\rho\psi = 0 \qquad \text{in } (0, 1) \qquad (2)$$

with, of course,

$$\psi(0) = 0. \qquad (3)$$

Therefore, from Equation (1), we are left with

$$p(1)\psi'(1)\phi(1) = 0$$

for *every* admissible ϕ. Since $p(x)$ is positive for $0 \leqslant x \leqslant 1$, this is possible only if, additionally,

$$\psi'(1) = 0. \qquad (4)$$

Thus the minimizing function ψ is an eigenfunction of the system (2), (3), (4) with corresponding eigenvalue μ, the minimum value of the quotient.

For the second part, we consider the system

$$\psi'' + \lambda\psi = 0 \qquad \text{in } (0, 1),$$

$$\psi(0) = \psi'(1) = 0.$$

For this system, the result just proved shows that for any smooth function ϕ satisfying $\phi(0) = 0$,

$$\frac{\int_0^1 \phi'^2}{\int_0^1 \phi^2} \geqslant \lambda_1$$

where λ_1 is the smallest eigenvalue. It is easy to confirm that the smallest eigenvalue is $\frac{1}{4}\pi^2$ (see *W: page 68*), and the result follows.

Solution to SAQ 11

We can rewrite the differential equation in two ways,

$$\frac{d^2}{dr^2}(ru) + \lambda ru = 0 \tag{1}$$

or

$$\frac{d}{dr}\left(r^2 \frac{du}{dr}\right) + \lambda r^2 u = 0. \tag{2}$$

(a) The first way enables us to get all the solutions: these are given by

$$u(r) = A\frac{\cos \lambda^{\frac{1}{2}} r}{r} + B\frac{\sin \lambda^{\frac{1}{2}} r}{r}.$$

The condition "$u(0)$ is finite" gives $A = 0$. The condition "$u(a) = 0$" gives the eigenvalues

$$\lambda_n = \frac{n^2\pi^2}{a^2} \qquad n = 1, 2, \ldots .$$

The solutions (eigenfunctions) are therefore given by

$$u_n(r) = \frac{\sin(n\pi r/a)}{r} \qquad n = 1, 2, \ldots .$$

(b) The second form, Equation (2), is the self-adjoint form of the given equation, and leads to the minimization problem:

minimize

$$\frac{\int_0^a r^2[\phi'(r)]^2\, dr}{\int_0^a r^2[\phi(r)]^2\, dr}$$

over all continuous and piecewise continuously differentiable functions ϕ on $[0, a]$ (in particular, finite at $r = 0$), satisfying $\phi(a) = 0$.

Solution to SAQ 12

(a) $R[\phi] = \dfrac{E(\phi, \phi)}{H(\phi, \phi)}$.

(b) We confirm the inner product axioms one by one.

H is

symmetric: Clearly $H(v, w) = H(w, v)$.

linear: $H(u, \alpha v + \beta w) = \int_0^1 \rho u(\alpha v + \beta w)$

$= \alpha H(u, v) + \beta H(u, w)$

for all $\alpha, \beta \in R$.

positive-definite: $H(u, u) > 0$ unless $u \equiv 0$ (since $\rho > 0$),

and

$H(u, u) = 0$ when $u \equiv 0$.

(c) E is

symmetric: Clearly $E(v, w) = E(w, v)$.

linear: $E(u, \alpha v + \beta w) = \int_0^1 [\rho u'(\alpha v' + \beta w') + qu(\alpha v + \beta w)]$

$= \alpha E(u, v) + \beta E(u, w)$.

positive-definite: $E(u, u) = \int_0^1 [pu'^2 + qu^2]$, which is greater than zero for every nonzero u provided that $q > 0$. Also when $u \equiv 0$, $E(u, u) = 0$.

(d) For a minimum, there exists an admissible function ψ such that

$$\frac{E(\psi + \tau\chi, \psi + \tau\chi)}{H(\psi + \tau\chi, \psi + \tau\chi)} \geqslant \frac{E(\psi, \psi)}{H(\psi, \psi)} = \mu, \text{ say,} \tag{1}$$

for all real τ and for all χ for which $\chi(0) = \chi(1) = 0$. Given any such function χ, the expression on the left attains its minimum when $\tau = 0$. The expansion of this term (using the symmetry and bilinearity of E and H) is

$$\frac{E(\psi, \psi) + 2\tau E(\psi, \chi) + \tau^2 E(\chi, \chi)}{H(\psi, \psi) + 2\tau H(\psi, \chi) + \tau^2 E(\chi, \chi)},$$

and the derivative at $\tau = 0$ is found to be

$$\frac{H(\psi, \psi)2E(\psi, \chi) - E(\psi, \psi)2H(\psi, \chi)}{[H(\psi, \psi)]^2}.$$

This is zero when

$$E(\psi, \chi) - \mu H(\psi, \chi) = 0,$$

using (1), and this must be true for every χ. Substitution of the integral representations of E and H gives Equation (36.9) in W again, and hence Equation (36.10).

(e) Let ϕ_B be a solution to Problem B. We wish to prove that it must also be a solution of Problem A. We make the hypothesis that ϕ_B is *not* a solution of Problem A, and show that this leads to a contradiction.

Let ϕ_A be any solution of Problem A. Then it is seen easily that any multiple of ϕ_A is also a solution of Problem A. We now define

$$\Phi_A = a\phi_A,$$

where $a = [H(\phi_A, \phi_A)]^{-\frac{1}{2}}$; Φ_A is therefore also a solution of Problem A, with the property

$$H(\Phi_A, \Phi_A) = a^2 H(\phi_A, \phi_A) = 1; \tag{2}$$

Φ_A is therefore an admissible function for Problem B. Now,

$$\frac{E(\phi_B, \phi_B)}{H(\phi_B, \phi_B)} > \frac{E(\Phi_A, \Phi_A)}{H(\Phi_A, \Phi_A)}, \tag{3}$$

since ϕ_B is, by hypothesis, not a solution of Problem A. But

$$H(\phi_B, \phi_B) = 1$$

by definition, and in view of Equation (2), (3) becomes

$$E(\phi_B, \phi_B) > E(\Phi_A, \Phi_A).$$

This contradicts the fact that ϕ_B is a solution to Problem B, and therefore our ϕ_B must be a solution to Problem A.

Solution to SAQ 13

We require the eigenvalues and eigenfunctions of the system

$$u'' + \lambda u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u(1) = 0.$$

These are given by $\lambda_n = n^2\pi^2$ $(n = 1, 2, \ldots)$ and $u_n(x) = \sin n\pi x$. Then, by (36.14),

$$G(x, y) = \sum_1^\infty \frac{\sin n\pi x \sin n\pi y}{n^2\pi^2 \int_0^1 \sin^2 n\pi\xi \, d\xi}$$

$$= \frac{2}{\pi^2} \sum_1^\infty \frac{\sin n\pi x \sin n\pi y}{n^2}.$$

Solution to SAQ 14

The functions given are the eigenfunctions of the system

$$u'' + \lambda u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u(1) = 0,$$

which is of the type considered in this section.

Solution to SAQ 15

We have

$$\lambda_k < \lambda_{k+1},$$

and u_{k+1}, u_k have respectively exactly k and $k - 1$ zeros in the open interval $(0, 1)$, and are also zero at the end points. By the Separation Theorem (with $\bar{\lambda} = \lambda_k$, $\lambda = \lambda_{k+1}$, $v = u_k$, $u = u_{k+1}$), there lies at least one zero of u_{k+1} on the open interval between each two successive zeros of u_k. Since there are k such intervals, the k zeros of u_{k+1} on $(0, 1)$ must be those on the open intervals, exactly one on each interval.

Solution to SAQ 16

Apply the Separation Theorem with $\bar{\lambda} = \lambda$ and (a) $u = u^*$, $v = u^{**}$, (b) $u = u^{**}$, $v = u^*$. The result follows since u^* and u^{**} are linearly independent.

Solution to SAQ 17

The eigenfunctions are displayed on *W: page 170, line 18*. The zeros x_m, say, of X_n $(n = 1, 2, \ldots)$ occur when

$$n\pi \frac{\log(1 + x_m)}{\log 2} = m\pi,$$

m being any integer, and $0 < x_m < 1$; i.e., when

$$x_m = 2^{\frac{m}{n}} - 1$$

for some integer m.

Now, since $n > 0$,

$$2^{\frac{m}{n}} - 1 > 0 \Leftrightarrow m > 0$$

and

$$2^{\frac{m}{n}} - 1 < 1 \Leftrightarrow m < n.$$

Therefore $0 < x_m < 1 \Leftrightarrow 0 < m < n$, so there are exactly $n - 1$ zeros, given by $m = 1, 2, \ldots, n - 1$.

For the second part the zeros of X_{n+1} are given by $2^{\frac{p}{n+1}} - 1$ $(p = 1, 2, \ldots, n)$. Therefore we need to show that

$$0 < 2^{\frac{1}{n+1}} - 1 < 2^{\frac{1}{n}} - 1 < 2^{\frac{2}{n+1}} - 1 < 2^{\frac{2}{n}} - 1 < \cdots < 2^{\frac{n}{n+1}} - 1 < 1,$$

which is obvious.

Solution to SAQ 18

Consider the boundary value problems

$$(pu')' - qu + \lambda \rho u = 0 \qquad \text{in } (0, 1),$$

$$u(0) = u(1) = 0, \tag{1}$$

$$p(x) = 1 + x^2,\ q(x) = x,\ \rho(x) = 1 + x^2,$$

and

$$(\bar{p}\bar{u}')' - \bar{q}\bar{u} + \bar{\lambda}\bar{\rho}\bar{u} = 0 \qquad \text{in } (0, 1),$$
$$\bar{u}(0) = \bar{u}(1) = 0. \tag{2}$$

Firstly, let

$$\bar{p} \equiv 1, \ \bar{q} \equiv 0, \ \bar{\rho} \equiv 2;$$

obviously $p \geqslant \bar{p}$, $q \geqslant \bar{q}$ and $\rho \leqslant \bar{\rho}$, and the two problems are not identical. By the Monotonicity Theorem we find that the eigenvalues of the two problems are related by

$$\bar{\lambda}_k < \lambda_k. \tag{3}$$

Problem (2) becomes

$$\bar{u}'' + 2\bar{\lambda}\bar{u} = 0 \qquad \text{in } (0, 1),$$
$$\bar{u}(0) = \bar{u}(1) = 0,$$

with eigenvalues

$$\bar{\lambda}_k = \tfrac{1}{2}k^2\pi^2.$$

Therefore, by (3),

$$\tfrac{1}{2}k^2\pi^2 < \lambda_k. \tag{4}$$

Similarly, let us compare with (1) the problem (2) with $\bar{p} \equiv 2$, $\bar{q} \equiv 1$, $\bar{\rho} \equiv 1$; then

$$2\bar{u}'' - \bar{u} + \bar{\lambda}\bar{u} = 0 \qquad \text{in } (0, 1),$$

or

$$\bar{u}'' + \tfrac{1}{2}(\bar{\lambda} - 1)\bar{u} = 0 \qquad \text{in } (0, 1),$$

and

$$\bar{u}(0) = \bar{u}(1) = 0.$$

The eigenvalues are given by

$$\tfrac{1}{2}(\bar{\lambda}_k - 1) = k^2\pi^2$$

or

$$\bar{\lambda}_k = 2k^2\pi^2 + 1.$$

By the Monotonicity Theorem, $2k^2\pi^2 + 1 > \lambda_k$. Therefore, from (4),

$$\tfrac{1}{2}k^2\pi^2 < \lambda_k < 2k^2\pi^2 + 1.$$

Solution to SAQ 19

(a) We proceed as in *W: page 172*. Choose the trial function

$$\phi(x) = \begin{cases} 0 & \alpha \leqslant x \leqslant \bar{\alpha}, \\ \bar{v}_1(x) & \bar{\alpha} \leqslant x \leqslant \bar{\beta}, \\ \bar{v}_1(\bar{\beta}) & \bar{\beta} \leqslant x \leqslant \beta, \end{cases}$$

where $\bar{v}_1$ is the first eigenfunction of

$$(\bar{p}\bar{v}')' - \bar{q}\bar{v} + \mu\bar{\rho}\bar{v} = 0 \qquad \text{in } (\bar{\alpha}, \bar{\beta}),$$
$$\bar{v}(\bar{\alpha}) = \bar{p}(\bar{\beta})\bar{v}'(\bar{\beta}) + \bar{b}\bar{v}(\bar{\beta}) = 0.$$

Clearly, ϕ is continuous and piecewise smooth on $[\alpha, \beta]$. We find that provided $q \equiv 0$,

$$\frac{\int_\alpha^\beta [p\phi'^2 + q\phi^2] + b[\phi(\beta)]^2}{\int_\alpha^\beta \rho\phi^2} \leqslant \frac{\int_\alpha^\beta [\bar{p}\bar{v}_1'^2 + \bar{q}\bar{v}_1^2] + \bar{b}[\bar{v}_1(\bar{\beta})]^2}{\int_\alpha^\beta \bar{\rho}\bar{v}_1^2} = \bar{\mu}_1,$$

since $\phi(\beta) = \bar{v}_1(\bar{\beta})$, $\bar{p} \geqslant p$, $\bar{q} \geqslant q$, $\bar{b} \geqslant b$ and $\bar{\rho} \leqslant \rho$. The proof then continues as in the text.

(b) Compare the given system with the system

$$v'' + \mu 2v = 0 \qquad \text{in } (0, 1),$$

$$v(0) = v'(1) = 0,$$

where we have decreased p and increased ρ. The lowest eigenvalue is given by

$$2\mu_1 = (\tfrac{1}{2}\pi)^2$$

or

$$\mu_1 = \tfrac{1}{8}\pi^2.$$

Therefore

$$\lambda_1 > \tfrac{1}{8}\pi^2,$$

by part (a) of the question.

Solution to SAQ 20

We proceed as in *W: pages 173 and 174*. Suppose that u_k has l zeros, at $\alpha = x_0$, $x_1, \ldots, x_{l-1} < \beta$. For $m = 1, 2, \ldots, l$ let

$$u^{(m)}(x) = \begin{cases} u_k(x) & x_{m-1} \leqslant x \leqslant x_m, \\ 0 & \text{elsewhere,} \end{cases}$$

where $x_l = \beta$, and let

$$\phi(x) = \sum_{m=1}^{l} c_m u^{(m)}(x), \tag{1}$$

where $c_1, \ldots, c_l$ are any constants. Then, as in note (ii) of Section 13.3.2,

$$\begin{aligned} &\int_\alpha^\beta [p\phi'^2 + q\phi^2] + b[\phi(\beta)]^2 \\ &= \sum_{m=1}^{l} c_m^2 \int_{x_{m-1}}^{x_m} [pu_k'^2 + qu_k^2] + b \sum_{m=1}^{l} c_m^2 [u^{(m)}(\beta)]^2 \\ &= \sum_{m=1}^{l} c_m^2 \left\{ \Big[pu_k u_k' \Big]_{x_{m-1}}^{x_m} - \int_{x_{m-1}}^{x_m} u_k[(pu_k')' - qu_k] \right\} + bc_l^2[u_k(\beta)]^2 \\ &= p(\beta)c_l^2 u_k(\beta)u_k'(\beta) - \sum_{m=1}^{l} c_m^2 \int_{x_{m-1}}^{x_m} u_k[(pu_k')' - qu_k] + bc_l^2[u_k(\beta)]^2 \end{aligned}$$

(noting that, this time, $u_k(\beta) \neq 0$)

$$= c_l^2 u_k(\beta)[p(\beta)u_k'(\beta) + bu_k(\beta)] - \sum_{m=1}^{l} c_m^2 \int_{x_{m-1}}^{x_m} u_k[(pu_k')' - qu_k].$$

The first term is zero, by the boundary condition at β; therefore, using the differential equation, this expression becomes

$$\lambda_k \sum_{m=1}^{l} c_m^2 \int_{x_{m-1}}^{x_m} \rho u_k^2 = \lambda_k \int_\alpha^\beta \rho\phi^2.$$

Hence,

$$\frac{\displaystyle\int_\alpha^\beta [p\phi'^2 + q\phi^2] + b[\phi(\beta)]^2}{\displaystyle\int_\alpha^\beta \rho\phi^2} = \lambda_k, \tag{2}$$

where ϕ is defined as in (1), and we have not so far restricted the coefficients $\{c_m\}$ in any way. The l constants $\{c_m\}$ may clearly be chosen to satisfy the $l-1$ linear equations

$$\int_\alpha^\beta \rho\phi u_r = 0 \qquad r = 1, 2, \ldots, l-1; \tag{3}$$

that is, we can arrange for ϕ to be orthogonal to the first $l-1$ eigenfunctions of the problem, without disturbing (2).

Now we know that if we minimize the quotient in (2) over *all* admissible functions ϕ which also satisfy Equations (3), we obtain a minimum equal to the lth eigenvalue λ_l. Therefore

$$\lambda_l \leqslant \lambda_k,$$

i.e.

$$l \leqslant k,$$

and so the number of zeros of u_k on (α, β), which is $l-1$, must be not greater than $k-1$.

Solution to SAQ 21

Let v be any solution of the equation

$$(pv')' - qv + \mu\rho v = 0,$$

with

$$v(\alpha) = 0$$

and let u satisfy

$$(pu')' - qu + \lambda\rho u = 0,$$

with

$$u(\alpha) = u(\beta) = 0.$$

Let l be an integer >1, and suppose that

$$\lambda_{l-1} < \mu < \lambda_l. \tag{1}$$

We shall show that v has exactly $l-1$ zeros on (α, β).

By the Separation Theorem, since $\lambda_{l-1} < \mu$, there is at least one zero of v between every consecutive pair of zeros of u_{l-1}. But by the Oscillation Theorem, there are exactly l zeros of u_{l-1} on $\alpha \leqslant x \leqslant \beta$, the end points being zeros. Therefore, if m is the number of zeros of v on (α, β),

$$m \geqslant l-1. \tag{2}$$

Again, by the Separation Theorem, since $\mu < \lambda_l$, there is at least one zero of u_l between each consecutive pair of zeros of v; also, there are exactly $l-1$ zeros of u_l on (α, β). Therefore, since $v(\alpha) = 0$,

$$m \leqslant l-1. \tag{3}$$

(2) and (3) together give

$$l-1 \leqslant m \leqslant l-1,$$

that is to say,

$$m = l-1. \tag{4}$$

Now let μ_k be an eigenvalue of the system (38.6). It cannot be equal to λ_l for any l, since in that case v_k would be an eigenfunction of both problems and satisfy both

$$v_k(\beta) = 0$$

and

$$p(\beta)v_k'(\beta) + bv_k(\beta) = 0.$$

That is to say,

$$v_k(\beta) = v_k'(\beta) = 0.$$

But by the Existence and Uniqueness Theorem this implies that $v_k \equiv 0$. Therefore μ_k must satisfy

$$\lambda_{l-1} < \mu_k < \lambda_l$$

for some l. It follows from (1) and (4), and the fact that v_k has exactly $k-1$ zeros on (α, β), that $l = k$.

Solution to SAQ 22

We compare the systems

$$v'' - q_{\max} v + \mu v = 0, \qquad v(0) = v(1) = 0; \tag{1}$$

$$u'' - qu + \lambda u = 0, \qquad u(0) = u(1) = 0;$$

$$w'' - q_{\min} w + \nu w = 0, \qquad w(0) = w(1) = 0; \tag{2}$$

where $q_{\max}$, $q_{\min}$ are the maximum and minimum values of q on $[0, 1]$.

The Monotonicity Theorem shows that, for each k,

$$\nu_k \leqslant \lambda_k \leqslant \mu_k, \tag{3}$$

with equality if and only if q is a constant function. But the eigenvalues of (1) are given by

$$\mu_k - q_{\max} = k^2\pi^2,$$

and those of (2) by

$$\nu_k - q_{\min} = k^2\pi^2.$$

Therefore, by (3),

$$k^2\pi^2 + q_{\min} \leqslant \lambda_k \leqslant k^2\pi^2 + q_{\max}$$

or

$$\pi^2 + \frac{q_{\min}}{k^2} \leqslant \frac{\lambda_k}{k^2} \leqslant \pi^2 + \frac{q_{\max}}{k^2}.$$

Now take the limit as $k \rightsquigarrow \infty$.

Solution to SAQ 23

(a) The eigenvalues of the systems with $\rho_{\max}$ or $\rho_{\min}$ in place of ρ are given by

$$\frac{k^2\pi^2}{(\beta-\alpha)^2\rho_{\max}}, \qquad \frac{k^2\pi^2}{(\beta-\alpha)^2\rho_{\min}}.$$

Since

$$\rho_{\min} \leqslant \rho(x) \leqslant \rho_{\max},$$

the result follows from the Monotonicity Theorem.

(b) The kth eigenfunction u_k has exactly $k-1$ zeros besides the two end points. On each interval (x_i, x_{i+1}), u_k must be the *first* eigenfunction for this interval, with the corresponding eigenvalue equal to λ_k. The result (a) applied *to this interval* therefore gives

$$\frac{\pi^2}{(x_{i+1}-x_i)^2\rho_{i,\max}} \leqslant \lambda_k \leqslant \frac{\pi^2}{(x_{i+1}-x_i)^2\rho_{i,\min}}.$$

It follows (after taking the square root) that

$$\rho_{i,\min}^{\frac{1}{2}}(x_{i+1}-x_i) \leqslant \frac{\pi}{\lambda_k^{\frac{1}{2}}} \leqslant \rho_{i,\max}^{\frac{1}{2}}(x_{i+1}-x_i),$$

for $i = 0, 1, \ldots, k-1$.

(c) Sum the inequalities obtained in (b), and put

$$x_{i+1} - x_i = \delta_i,$$

so that

$$\sum_{i=0}^{k-1} \rho_{i,\min}^{\frac{1}{2}} \delta_i \leqslant \frac{k\pi}{\lambda_k^{\frac{1}{2}}} \leqslant \sum_{i=0}^{k-1} \rho_{i,\max}^{\frac{1}{2}} \delta_i.$$

Now assume (we have not proved this) that

$$\lim_{k \sim \infty} M_k = 0,$$

so that all the intervals (x_i, x_{i+1}) tend to zero in length. Then we obtain in the limit

$$\int_\alpha^\beta \rho^{\frac{1}{2}}(x)\, dx = \lim_{k \sim \infty} \frac{k\pi}{\lambda_k^{\frac{1}{2}}},$$

by the definition of the integral. This may be rewritten, after squaring, as

$$\lim_{k \sim \infty} \frac{\lambda_k}{k^2} = \frac{\pi^2}{\left[\int_\alpha^\beta \rho^{\frac{1}{2}}(x)\, dx\right]^2}, \tag{1}$$

which gives an estimate for the eigenvalues for large k.

The final result indicates that, for large k, the eigenvalues tend to those of a certain simple harmonic problem

$$u'' + \lambda \tilde{\rho} u = 0, \qquad u(\alpha) = u(\beta) = 0,$$

where

$$\tilde{\rho} = \left[\int_\alpha^\beta \rho^{\frac{1}{2}}(x)\, dx\right]^2;$$

this is a kind of average value of ρ on the interval, though not, perhaps, the expected one.

As an example, consider the variable string problem of *W: Section 37*, with

$$\rho(x) = \frac{1}{(1+x)^2} \quad \text{and} \quad \lambda_k = \frac{k^2\pi^2}{(\log 2)^2} + \frac{1}{4}.$$

Equation (1) predicts

$$\lim_{k \sim \infty} \frac{\lambda_k}{k^2} = \frac{\pi^2}{\left[\int_0^1 \frac{dx}{(1+x)}\right]^2} = \frac{\pi^2}{(\log 2)^2},$$

which is clearly correct. k does not have to be very large for this estimate of the eigenvalues to be quite good; for example, the error for $k = 2$ is 0.3%.

Unit 14 Bessel Functions

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differential Equations* (Oxford, 1971).

H. F. Weinberger, *A First Course in Partial Differential Equations* (Xerox, 1965).

It is essential to have these books; the course is based on them and will not make sense without them. They are referred to in the text as *S* and *W* respectively.

Unit 14 is based on *W*: Chapter VII, Sections 40 to 42.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics.* References to Open University courses in mathematics take the form:

> *Unit M100 13, Integration II* for the Mathematics Foundation Course,
> *Unit M201 23, The Wave Equation* for the Linear Mathematics Course.

Bibliography

M. Abramowitz and Irene A. Stegun (eds.) *Handbook of Mathematical Functions* (Dover, 1965).

E. Jahnke and F. Emde, *Tables of Functions* (Dover, 1945).

These are reference works which list the definitions and properties of numerous special functions, including Bessel functions; however, the results are not proved.

D. L. Kreider, R. G. Kuller, D. R. Ostberg and F. W. Perkins, *An Introduction to Linear Analysis* (Addison-Wesley, 1966).

You will find a useful discussion of Bessel functions in Chapter 15.

G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, first published 1922).

This classic gives a comprehensive treatment of the subject, well beyond the scope of this course. An account of the history of the subject appears in Chapter 1.

14.0 INTRODUCTION

This unit is concerned with *Bessel functions* and some typical boundary value problems in which they arise. Bessel functions arise as solutions of a certain second-order ordinary differential equation, known as *Bessel's equation*. The interest in such functions in a course on partial differential equations is due to the fact that Bessel's equation appears, for example, when we apply the separation of variables technique to time-dependent problems such as the two-dimensional wave equation and heat conduction equation in polar coordinates.

Bessel functions belong to a group of functions which go under the general name of *special functions*. If you want to get some idea of the extent of the subject of special functions, borrow a copy of *Handbook of Mathematical Functions* edited by M. Abramowitz and Irene A. Stegun.

Following the treatment in *W*, we look firstly at some properties of Bessel functions and then consider, by the use of illustrations, just how they arise in physical problems.

14.1 BESSEL FUNCTIONS OF THE FIRST KIND

14.1.1 Solution in Series

READ W: Section 40, page 179 to page 180, line 2.

Notes

(i) *W: page 179, lines 2 to 7*
W: Section 39 is not included in the course and consequently this reading passage is your first contact with this particular boundary value problem. Completeness of the eigenfunctions will be assumed.

(ii) *W: page 179, Equation (40.3)*
Bessel's equation (40.3) is often written as

$$t^2\frac{d^2u}{dt^2} + t\frac{du}{dt} + (t^2 - m^2)u = 0.$$

We assume that m is nonnegative.

(iii) *W: page 179, line 13*
The problem is to find constants $\alpha, a_0, a_1, a_2, \ldots$ which ensure that the series satisfies the differential equation. The condition $a_0 \neq 0$ is not a restriction but merely a way of stating that α is the first index in the series. Since the equation is of the second order we expect two such series solutions. A series of the form

$$\sum_{k=0}^{\infty} b_k t^k$$

is called a *power series*

(iv) *W: page 179, lines 14 to 16*
This note fills in some gaps in the manipulation. Term-by-term differentiation of the series gives

$$\frac{du}{dt}(t) = \sum_{n=0}^{\infty} (\alpha + n)a_n t^{\alpha+n-1}$$

so that

$$t\frac{du}{dt}(t) = \sum_{n=0}^{\infty} (\alpha + n)a_n t^{\alpha+n}.$$

Differentiating both sides again with respect to t, we find that

$$\frac{d}{dt}\left(t\frac{du}{dt}(t)\right) = \sum_{n=0}^{\infty} (\alpha + n)^2 a_n t^{\alpha+n-1}.$$

We now substitute the series for u and $d/dt(t\,du/dt)$ in Equation (40.3) and obtain

$$t^{\alpha-1}\left\{\sum_{n=0}^{\infty} (\alpha + n)^2 a_n t^n - m^2 \sum_{n=0}^{\infty} a_n t^n + \sum_{n=0}^{\infty} a_n t^{n+2}\right\} = 0.$$

The last series on the left-hand side may be rewritten as

$$\sum_{n=0}^{\infty} a_n t^{n+2} = \sum_{n=2}^{\infty} a_{n-2} t^n,$$

and *line 16* follows.

(v) *W: page 179, lines −11 to −7*
The choice $2^{-m}/m!$ for a_0 is conventional and the notation $J_m(t)$ is introduced with this value of a_0. Note that all the remaining coefficients are completely determined. Since $a_1 = 0$ all coefficients with odd suffixes must vanish by the recursion formula

$$a_n = -\frac{a_{n-2}}{(m+n)^2 - m^2} = -\frac{a_{n-2}}{(2m+n)n}.$$

For the other coefficients we have successively

$$a_2 = -\frac{a_0}{2(m+1)\,.\,2} = -\frac{2^{-m-2}}{(m+1)!},$$

$$a_4 = -\frac{a_2}{2(m+2)\,.\,4} = \frac{2^{-m-4}}{2(m+2)!},$$

and, by induction,

$$a_{2k} = -\frac{a_{2k-2}}{2(m+k)\,.\,2k} = \frac{(-1)^k 2^{-m-2k}}{k!(m+k)!}.$$

Equation (40.4) follows by substituting these coefficients into the series.

J_m is bounded at $t = 0$ because m is assumed to be nonnegative. It is of course the boundedness condition which forbids us the choice $\alpha = -m$.

(vi) *W: page 179, footnote*
If m is not an integer, the convention for a_0 becomes

$$a_0 = \frac{2^{-m}}{\Gamma(m+1)},$$

where the **Gamma function** Γ is defined, for $m > 0$, by

$$\Gamma(m) = \int_0^\infty e^{-x} x^{m-1}\,dx.$$

When m is a positive integer, $\Gamma(m+1) = m!$ (See SAQ 3.)

(vii) *W: page 179, line −5*
The formal construction of a series solution of a differential equation does not ensure that the series *converges* for all (or any) values of t. Before we can really define $J_m(t)$ by Equation (40.4) we must examine for what values of t the series converges. The **ratio test** is a test for convergence. It states that if, in the series $\sum_{k=0}^{\infty} c_k$, $c_k \neq 0$ and

$$\lim_{k\sim\infty}\left|\frac{c_{k+1}}{c_k}\right| < 1$$

then the series $\sum_{k=0}^{\infty} |c_k|$ converges. Now a series $\sum c_k$ for which $\sum |c_k|$ converges is said to be *absolutely convergent*, and there is a result which states than an absolutely convergent series is convergent.*

In the series (40.4) put

$$c_k = \frac{(-1)^k(\frac{1}{2}t)^{m+2k}}{k!(m+k)!},$$

whence, for $t \neq 0$,

$$\lim_{k\sim\infty}\left|\frac{c_{k+1}}{c_k}\right| = \lim_{k\sim\infty}\frac{\frac{1}{4}|t|^2}{(k+1)(m+k+1)} = 0$$

for all $t \neq 0$. Thus, using the ratio test, and noting that the series obviously converges for $t = 0$, we see that the series converges for all t. For nonnegative values of m the Bessel function J_m is defined by the power series (40.4).

General Comment

You may wonder what happens if we consider the unbounded solution of Bessel's equation. Put $\alpha = -m$ and carry the steps through formally. Then apparently

$$J_{-m}(t) = \sum_{k=0}^{\infty}\frac{(-1)^k(\frac{1}{2}t)^{-m+2k}}{k!\Gamma(-m+k+1)}$$

*These results may be found in M. Spivak, *Calculus* (Benjamin, 1973), pp. 393 and 397.

for any nonnegative m. But the terms of the power series are not all defined; for example, for $k = 0$ and $m = \frac{5}{2}$,

$$\Gamma(-m + k + 1) = \Gamma(-\tfrac{3}{2})$$

and the Gamma function cannot be defined by the integral formula in note (vi) since the integral diverges. However, provided $0 < m < 1$ we get a satisfactory definition for J_{-m}. (In fact there is a different definition of the Gamma function which applies in the extended domain containing negative arguments, although we do not intend to go into this here.) What we do is look into the case $\alpha = -m$ by using the recursion formula

$$(n - 2m)na_n = -a_{n-2}.$$

Since $a_1 = 0$ again all the odd coefficients vanish. If m is not an integer then we can make some choice of a_0 and obtain another solution, independent of J_m, which will be unbounded as $t \rightsquigarrow 0$. If m is an integer, it follows from the recursion formula that $a_{2m-2} = 0$, and by successive application of the formula we conclude that

$$a_{2r} = 0 \qquad r = 0, 1, \ldots, m - 1.$$

Thus, if m is an integer, we do *not* obtain a further independent solution. (Although $a_0 = 0$ is incompatible with the defining condition $a_0 \neq 0$ we could investigate this case further. The first nonzero coefficient would then be a_{2m}. Assign the value $(-1)^m 2^{-m}/m!$ to a_{2m}. It then follows that we could define

$$J_{-m}(t) = \sum_{k=m}^{\infty} \frac{(-1)^k(\frac{1}{2}t)^{-m+2k}}{k!(k-m)!}.$$

A change of the dummy variable from k to j by $k = j + m$ now gives

$$J_{-m}(t) = \sum_{j=0}^{\infty} \frac{(-1)^{j+m}(\frac{1}{2}t)^{m+2j}}{(j+m)!j!} = (-1)^m J_m(t),$$

confirming that no new solution has been obtained.)

SAQ 1

Show that

$$\frac{dJ_0(t)}{dt} = -J_1(t).$$

(Solution on p. 20.)

SAQ 2

W: page 181, Exercise 2

(Solution on p. 20.)

SAQ 3

(a) Using the definition of the Gamma function show that

$$\Gamma(1) = 1, \quad \Gamma(2) = 1, \quad \Gamma(3) = 2.$$

(b) Show that

$$\Gamma(m + 1) = m\Gamma(m)$$

for all $m > 0$. Deduce that if m is a positive integer $\Gamma(m + 1) = m!$.

(Solution on p. 20.)

SAQ 4

Given the result

$$\int_0^\infty e^{-t^2}\,dt = \tfrac{1}{2}\pi^{\frac{1}{2}},$$

deduce that $\Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$.

(Solution on p. 20.)

SAQ 5

Supply the details to establish Equations (40.5) and (40.6) in *W: pages 179 and 180*.

(Solution on p. 21.)

SAQ 6

Verify that

$$J_0(t) = \frac{1}{\pi}\int_0^\pi \cos(t\sin\theta)\,d\theta.$$

(Solution on p. 21.)

14.1.2 Zeros of Bessel Functions

READ W: page 180, line 3 to page 181, line 17 (the end of the section).

Notes

(i) *W: page 180, line 11*
That $\lambda_k^{(m)}$ increases with m follows from the Monotonicity Theorem on *W: page 174*.

(ii) *W: page 180, lines 13 and 14*
Since J_m satisfies Bessel's equation it is certainly continuous for $t > 0$. Therefore, between any two successive zeros the function $t^{-m}J_m$ must have at least one stationary value. At this point $d[t^{-m}J_m]/dt$ must vanish; thus by (40.5), J_{m+1} must have at least one zero between successive positive zeros of J_m. We prove, by contradiction, that J_{m+1} has just one zero in this interval. Suppose J_{m+1} had more than one zero in $(j_k^{(m)}, j_{k+1}^{(m)})$. Then

$$\frac{d}{dt}[t^{m+1}J_{m+1}(t)] = 0$$

between two such zeros, and from (40.6) with m replaced by $m + 1$ it would follow that $J_m(t) = 0$ for some t in $(j_k^{(m)}, j_{k+1}^{(m)})$. This would contradict the assumption that the original zeros of J_m which we chose were successive. Thus the zeros of J_m and J_{m+1} are interlaced.

(iii) *W: page 180, Equation (40.9)*
This problem is not equivalent to the original eigenvalue problem (40.1) and (40.2). Solutions for which $w(0) = 0$ do not necessarily satisfy the conditions

u bounded

$\lim xu' = 0$ as $x \rightsquigarrow 0$.

However, we now know the properties of $J_m(\sqrt{\lambda}x)$ as $x \rightsquigarrow 0$.

(iv) *W: page 181, lines 4 to 6*
The notation

$$f(x) \sim g(x)$$

is often used to mean that

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$

In this case we often say that $g(x)$ is an *asymptotic form* for $f(x)$.

(v) *W: page 181, lines 8 and 9*
In the terminology of *W: page 162, line −1* the eigenfunctions are

$$u_k(x) = J_m(\sqrt{\lambda_k^{(m)}}x)$$

and the weight function is $\rho(x) \equiv x$. The integral

$$\int_0^1 xJ_m(\sqrt{\lambda_k^{(m)}}x)^2\,dx$$

appears in the denominator of the Fourier coefficient c_k.

(vi) *W: page 181, line 11*
In the right-hand side of the equation $J_m^2(t)$ means $[J_m(t)]^2$.

The following example shows you how to construct the **Fourier-Bessel series** (as it is known) for a given function.

Example

Find the expansion of 1 in the interval (0, 1) in terms of the eigenfunctions $J_0(\sqrt{\lambda_k^{(0)}}x)$.

We require the coefficients A_k in the series

$$1 = \sum_{k=1}^{\infty} A_k J_0(\sqrt{\lambda_k^{(0)}}x) \qquad x \in (0, 1).$$

We multiply both sides by $xJ_0(\sqrt{\lambda_k^{(0)}}x)$ and integrate from 0 to 1. By the orthogonality property of the eigenfunctions we are left with one term on the right-hand side. Thus

$$\int_0^1 xJ_0(\sqrt{\lambda_n^{(0)}}x)\,dx = A_n \int_0^1 x[J_0(\sqrt{\lambda_n^{(0)}}x)]^2\,dx$$
$$= \tfrac{1}{2}A_n[J_1(\sqrt{\lambda_n^{(0)}})]^2$$

by *W: page 181, line 17*. The integral on the left becomes

$$\int_0^1 xJ_0(\sqrt{\lambda_n^{(0)}}x)\,dx = \frac{1}{\lambda_n^{(0)}} \int_0^{\sqrt{\lambda_n^{(0)}}} tJ_0(t)\,dt$$
$$= \frac{1}{\lambda_n^{(0)}} \Big[tJ_1(t) \Big]_0^{\sqrt{\lambda_n^{(0)}}} \qquad \text{by Equation (40.6)}$$
$$= \frac{1}{\sqrt{\lambda_n^{(0)}}} J_1(\sqrt{\lambda_n^{(0)}}).$$

Hence

$$A_n = \frac{2}{\sqrt{\lambda_n^{(0)}}J_1(\sqrt{\lambda_n^{(0)}})}$$

and

$$1 = \sum_{k=1}^{\infty} \frac{2}{\sqrt{\lambda_k^{(0)}}J_1(\sqrt{\lambda_k^{(0)}})} J_0(\sqrt{\lambda_k^{(0)}}x) \qquad x \in (0, 1).$$

SAQ 7

W: page 181, Exercise 3.

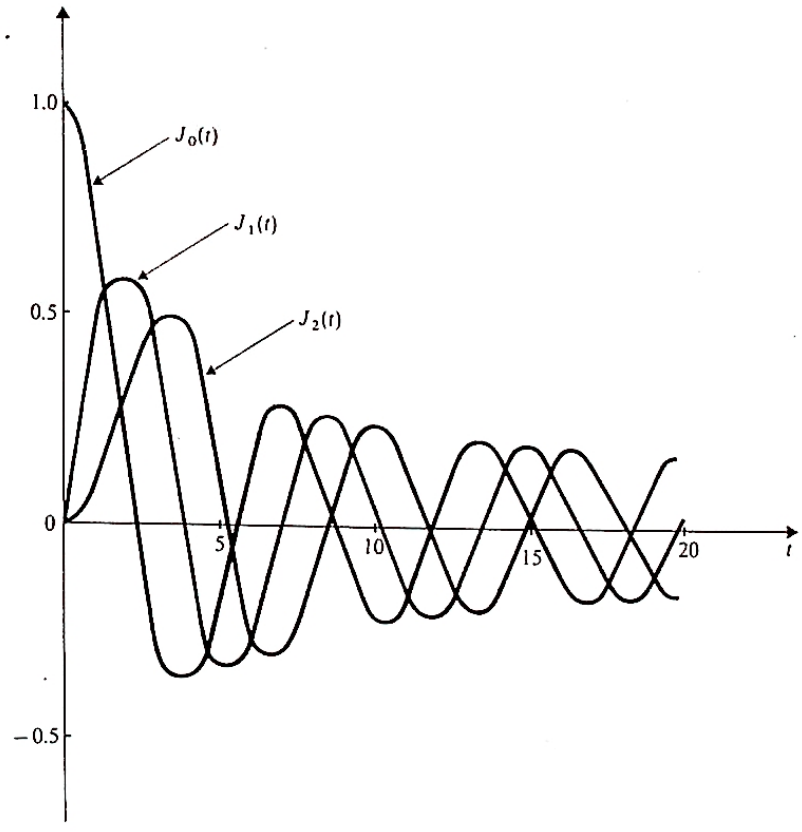
(Solution on p. 22.)

SAQ 8

W: page 181, Exercise 4.

(Solution on p. 22.)

Extensive tables of Bessel functions for different orders exist (see Bibliography). The following sketch shows graphs of $J_0(t)$, $J_1(t)$ and $J_2(t)$ in the interval $0 \leqslant t \leqslant 20$, and their oscillatory character (cf. the Oscillation Theorem) is clearly visible.



The table below gives the first few zeros of $J_0(t)$, $J_1(t)$ and $J_2(t)$. Here $j_k^{(m)}$ represents the kth positive zero of $J_m(t)$.

k	$j_k^{(0)}$	$j_k^{(1)}$	$j_k^{(2)}$
1	2.405	3.832	5.136
2	5.520	7.016	8.417
3	8.654	10.173	11.620
4	11.792	13.324	14.796
5	14.931	16.471	17.960

We shall see in Section 14.2 that the Bessel functions of integral order arise as solutions of an equation which is derived by modelling oscillatory systems in two dimensions. This is the physical background to the oscillatory nature of the J_m. You might wonder whether these can be related in a simple way to the more elementary trigonometric functions which have the same oscillatory property. In general this is not the case; there is however one special set of Bessel functions for which there is a simple relationship with the trigonometric functions. This is the set of Bessel functions of half-integral order, i.e.

$$J_{m+\frac{1}{2}} \qquad m = 0, 1, 2, \ldots .$$

(The functions

$$r \longmapsto \left(\frac{\pi}{2r}\right)^{\frac{1}{2}} J_{m+\frac{1}{2}}(r)$$

are known as *spherical Bessel functions*; they arise in the solution to the wave equation in spherical polar coordinates. We shall not however give a derivation here.)

We now obtain the relationships by deriving firstly the one for $J_{\frac{1}{2}}$, and then using *W: page 179, Equation (40.5)* to determine the others. $J_{\frac{1}{2}}$ is the solution of

$$(tu')' - \frac{u}{4t} + tu = 0$$

which is bounded at $t = 0$.

Make the substitution

$$w(t) = u(t)t^{\frac{1}{2}};$$

then, as in Equation (40.9), w satisfies

$$w'' + w = 0,$$

$$w(0) = 0.$$

This has the independent solutions $\sin t$ and $\cos t$ and, since $w(0) = 0$, we must discard $\cos t$ and get $w(t) = A \sin t$ as the solution. Thus we can argue that

$$J_{\frac{1}{2}}(t) = \frac{1}{t^{\frac{1}{2}}} w(t) = \frac{A \sin t}{t^{\frac{1}{2}}}.$$

We can settle the value of A by taking the limit as $t \rightsquigarrow 0$:

$$\lim_{t \rightsquigarrow 0} \frac{J_{\frac{1}{2}}(t)}{t^{\frac{1}{2}}} = \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{3}{2})} \qquad \text{from Equation (40.4), suitably adjusted}$$

whilst

$$\lim_{t \rightsquigarrow 0} \frac{\sin t}{t} = 1.$$

Matching these leads to

$$A = \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{3}{2})} = \left(\frac{2}{\pi}\right)^{\frac{1}{2}}$$

since $\Gamma(\frac{3}{2}) = \frac{1}{2}\Gamma(\frac{1}{2})$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (see SAQ 3 and SAQ 4). Thus, finally,

$$J_{\frac{1}{2}}(t) = \left(\frac{2}{\pi t}\right)^{\frac{1}{2}} \sin t.$$

[You may also derive this result by writing out and reorganizing the series expansion *W: Equation (40.4)*, and recognizing the Taylor expansion of the sine function.]

To determine the formula for the other Bessel functions of half-integral order we write *W: Equation (40.5)* as

$$\frac{1}{t^{m+1}} J_{m+1}(t) = -\frac{1}{t} \frac{d}{dt}\left[\frac{1}{t^m} J_m(t)\right].$$

We may use this formula recursively to obtain, by induction,

$$\frac{1}{t^{m+n}} J_{m+n}(t) = \left(-\frac{1}{t} \frac{d}{dt}\right)^n \left[\frac{1}{t^m} J_m(t)\right].$$

In particular, for $m = \frac{1}{2}$ we have

$$\frac{1}{t^{n+\frac{1}{2}}} J_{n+\frac{1}{2}}(t) = \left(-\frac{1}{t} \frac{d}{dt}\right)^n \left[\frac{1}{t^{\frac{1}{2}}} J_{\frac{1}{2}}(t)\right]$$

or

$$J_{n+\frac{1}{2}}(t) = t^n \left(\frac{2t}{\pi}\right)^{\frac{1}{2}} \left(-\frac{1}{t} \frac{d}{dt}\right)^n \left(\frac{\sin t}{t}\right).$$

We have indicated in Section 14.1.1 that for $m > 0$ we could obtain another (independent) solution of Bessel's equation by the power series method, choosing $\alpha = -m$. Provided m is not an integer we obtain a new solution, independent of J_m, which is unbounded at the origin. For $m = \frac{1}{2}$ the series

$$J_{-\frac{1}{2}}(t) = \sum_{k=0}^{\infty} \frac{(-1)^k (\frac{1}{2}t)^{2k-\frac{1}{2}}}{k!\Gamma(k + \frac{1}{2})}$$

is obtained. For $m = n + \frac{1}{2}$ $(n = 1, 2, \ldots)$ the expression $\Gamma(-m + 1)$ is not defined. We can however *define* J_{-m} recursively using Equation (40.6). It is possible to show

that the function J_{-m} so constructed is a solution of Bessel's equation of order m and is not linearly dependent on J_m. In the next SAQ we ask you to derive an expression for $J_{-n-\frac{1}{2}}(t)$ when $n = 0, 1, 2, \ldots$.

SAQ 9

(a) Show that

$$J_{-\frac{1}{2}}(t) = \left(\frac{2}{\pi t}\right)^{\frac{1}{2}} \cos t.$$

HINT: $J_{-\frac{1}{2}}$ and $J_{\frac{1}{2}}$ are related by *W: page 180, Equation (40.6)*.

(b) Deduce that

$$J_{-n-\frac{1}{2}}(t) = t^n \left(\frac{2t}{\pi}\right)^{\frac{1}{2}} \left(\frac{1}{t}\frac{d}{dt}\right)^n \left(\frac{\cos t}{t}\right)$$

for $n = 0, 1, 2, \ldots$.

(Solution on p. 23.)

14.2 TIME-DEPENDENT PROBLEMS IN TWO DIMENSIONS

14.2.1 Vibration of a Circular Membrane

READ W: Section 41, page 182 top of page to line −7.

Note

(i) *W: page 182*
This is the membrane problem which first appeared in *W: page 48*. The membrane, which is stretched across a fixed plane circular wire, is given an initial transverse displacement $u(r, \theta, 0) = f(r, \theta)$ and released from rest. Subsequently, the membrane vibrates freely.

READ W: Section 42, pages 185 to 187, omitting page 186, lines 8 to 12.

SAQ 10

Show that the solution to the vibrating circular membrane problem (*W: page 182*) has the form (41.3).

(Solution on p. 24.)

SAQ 11

W: page 185, Exercise 1(b)

(Solution on p. 24.)

SAQ 12

A membrane is stretched across a circular frame of unit radius. The transverse displacement $u(\mathbf{r}, t)$ satisfies

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u.$$

The membrane is released from rest in the position $u(\mathbf{r}, 0) = \varepsilon(1 - r^2)$ where ε is small and r is the distance from the centre. Find the subsequent displacement of the membrane.

(You will need the result of SAQ 7.)

(Solution on p. 25.)

14.2.2 Sound Waves in a Tube

We begin with a generalization of the discussion in *Unit 1, The Wave Equation* in which the one-dimensional wave equation was derived for the propagation of sound in a gas. We show how the wave equation in two or three dimensions follows from the equations of motion, which are:

Equation of Continuity

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0, \tag{1}$$

which expresses conservation of mass; (See Appendix.)

Momentum Equation

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \mathbf{grad})\,\mathbf{v}\right) = -\mathbf{grad}\, p, \tag{2}$$

assuming the gas is subject to no external forces [this equation may be interpreted componentwise, i.e., for any fixed unit vector e

$$\rho \frac{dv_e}{dt} \equiv \rho\left(\frac{\partial v_e}{\partial t} + \mathbf{v} \cdot \mathbf{grad}\, v_e\right) = -\mathbf{e} \cdot \mathbf{grad}\, p \cdot$$

where $v_e = \mathbf{e} \cdot \mathbf{v}$ is the component of v in the direction of e]; (See Appendix.)

Equation of State

$$p = \alpha \rho^{\gamma}, \tag{3}$$

where α and γ are constants depending on the particular gas. (See *Unit 1*, Section 1.1.3.)

These equations provide relations between the pressure $p(\mathbf{r}, t)$, the density $\rho(\mathbf{r}, t)$ and the gas velocity $\mathbf{v}(\mathbf{r}, t)$.

Suppose now that a sound wave passes through a gas at rest with initial pressure $p(\mathbf{r}, 0) = p_0$ and density $\rho(\mathbf{r}, 0) = \rho_0$. Equation (3) must be satisfied in the initial state so that

$$\alpha = \frac{p_0}{\rho_0^{\gamma}}.$$

Put $\rho = \rho_0(1 + s)$, where we assume as part of the approximation that s is small. Thus, from (3)

$$p = \frac{p_0}{\rho_0^{\gamma}} \rho_0^{\gamma}(1 + s)^{\gamma} \simeq p_0(1 + \gamma s),$$

where we have retained just the first two terms in the binomial expansion of the right-hand side. The basis of the approximation is that s and v and their derivatives are small so that *second* degree terms containing them may be neglected. Thus Equations (1) and (2) become

$$\frac{\partial s}{\partial t} + \operatorname{div} \mathbf{v} = 0, \tag{4}$$

$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} = -\mathbf{grad}\, p \simeq -p_0 \gamma\, \mathbf{grad}\, s. \tag{5}$$

Assuming that ρ (and hence s) is twice continuously differentiable we have

$$\operatorname{div} \frac{\partial \mathbf{v}}{\partial t} = -\frac{\partial^2 s}{\partial t^2} \qquad \text{from (4)}$$

$$= -\frac{\gamma p_0}{\rho_0} \operatorname{div} \mathbf{grad}\, s. \qquad \text{from (5)}$$

We set $c^2 = \gamma p_0/\rho_0$ and obtain

$$\nabla^2 s - \frac{1}{c^2} \frac{\partial^2 s}{\partial t^2} = 0,$$

i.e. the density satisfies the two- or three-dimensional wave equation.

Since the gas starts from rest it may be shown that, as a result of Equation (5), there is a scalar field Φ such that

$$\mathbf{v} = -\mathbf{grad}\, \Phi.$$

Note that Φ is undetermined to the extent of an arbitrary function of time. We call Φ the **velocity potential.**

Rearranging (5) we find that

$$\mathbf{grad}\left(\frac{\gamma p_0}{\rho_0} s - \frac{\partial \Phi}{\partial t}\right) = 0, \tag{6}$$

whence

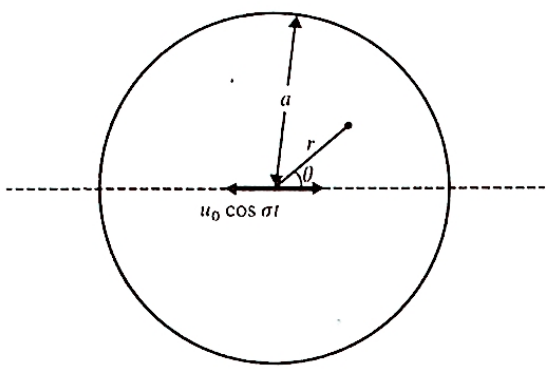
$$s = \frac{1}{c^2}\frac{\partial \Phi}{\partial t}.$$

An arbitrary function of time could appear from (6) but we absorb that into $\partial\Phi/\partial t$. Finally substituting the expressions for s and $\mathbf{v}$ into Equation (4) we get the wave equation

$$\frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = \text{div grad } \Phi \equiv \nabla^2 \Phi.$$

The number c is a constant which measures the rate of propagation of disturbances and for this reason is known as the *speed of sound* in the gas.

Suppose air is enclosed by a long hollow tube of circular cross-section with radius a. The tube as a whole vibrates rigidly perpendicular to its axis with velocity $u_0 \cos \sigma t$, that is, it oscillates harmonically. Let us find out how the air inside the tube vibrates.



For a long tube it is arguable that, away from the ends of the tube, the velocity potential is independent of distance along the tube and varies only with time t and position relative to the axis of the tube and the direction of vibration. Polar coordinates are suitable for this problem because the boundary is a circle in the cross-section. Thus Φ satisfies

$$\frac{\partial^2 \Phi}{\partial r^2} + \frac{1}{r}\frac{\partial \Phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \Phi}{\partial \theta^2} = \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} \qquad 0 < r < a,\ 0 < \theta < 2\pi,$$

$$\Phi|_{\theta=0} = \Phi|_{\theta=2\pi},$$

and Φ is bounded as $r \rightsquigarrow 0$. How does the air immediately adjacent to the wall behave? We take no account of viscosity; so we suppose that the air may slip freely *parallel* to the wall of the tube, but must have the same velocity *normal* to the wall as the wall itself. Let $\theta = 0$ be the direction of vibration. The normal component of the wall's velocity at points with angular coordinate θ is $u_0 \cos \sigma t \cos \theta$, whilst the normal velocity of the air is $-\partial\Phi/\partial r$. Thus

$$-\left.\frac{\partial \Phi}{\partial r}\right|_{r=a} = u_0 \cos \sigma t \cos \theta.$$

The boundary conditions suggest we try a solution of the form

$$\Phi = f(r) \cos \theta \cos \sigma t$$

where, after substitution into the wave equation, f satisfies

$$f''(r) + \frac{1}{r}f'(r) - \frac{1}{r^2}f(r) + k^2 f(r) = 0 \qquad 0 < r < a,$$

$$f'(a) = -u_0,$$

and $k = \sigma/c$. This is Bessel's equation of order 1. The solution which is bounded at $r = 0$ is

$$f(r) = AJ_1(kr).$$

The constant A is determined by the boundary condition

$$kAJ_1'(ka) = -u_0 \quad \Rightarrow \quad A = -\frac{u_0}{kJ_1'(ka)}.$$

The required solution is therefore

$$\Phi = -\frac{u_0 c}{\sigma}\frac{J_1(kr)}{J_1'(ka)}\cos\theta\cos\sigma t.$$

Since J_1 is an oscillating function it follows that J_1' also oscillates about zero; in other words J_1' has an unbounded number of zeros. If ka coincides with any one of these zeros then the solution breaks down. In the neighbourhood of these zeros the amplitude of the wave is very large: the linear theory of sound waves is then no longer appropriate and the assumptions made in the derivation of the wave equation must be re-examined.

14.2.3 Heat Conduction in a Bar

We complete this section by looking at an example in heat conduction.

Consider a long bar of circular cross-section and unit radius which has an initial temperature distribution which is radially symmetric and does not vary along the bar. Suppose the boundary temperature is maintained at a constant value, say zero. The bar is supposed to be long so that we can ignore end effects. What is the temperature distribution at time t over the cross-section of the bar? Since the temperature depends only on the radius r and time t, the heat conduction equation becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) = \frac{1}{k}\frac{\partial u}{\partial t} \qquad 0 < r < 1, t > 0,$$

where $u(r, t)$ represents the temperature. The boundary and initial conditions are respectively

$$u(1, t) = 0 \qquad t > 0$$

and

$$u(r, 0) = f(r) \qquad 0 \leqslant r < 1$$

where f is a prescribed function. We also require that $\lim_{t\to\infty} u(r, t) = 0$ and that $u(r, t)$ be bounded as $r \to 0$.

To apply the separation of variables method we look for solutions of the form $R(r)T(t)$. Thus

$$\frac{1}{rR}\frac{d}{dr}\left(r\frac{dR}{dr}\right) = \frac{1}{kT}\frac{dT}{dt} = -\lambda,$$

say. The functions R and T satisfy the ordinary differential equations

$$(rR')' + \lambda rR = 0 \qquad 0 < r < 1, \quad R(0) \text{ bounded}, \quad R(1) = 0,$$

and

$$T' + \lambda kT = 0 \qquad t > 0, \quad \lim_{t\to\infty} T(t) = 0.$$

The second problem has the solution

$$T(t) = e^{-\lambda kt}.$$

We require the solution to decrease with time and so we impose the condition $\lambda > 0$. The first equation is Bessel's equation of order zero with a solution $J_0(\sqrt{\lambda}r)$. The other independent solution of Bessel's equation has not been investigated when the order is an integer but it has a singularity at $r = 0$, whereas J_0 is bounded at $r = 0$.

The boundary condition $R(1) = 0$ is satisfied if we choose λ so that

$$J_0(\sqrt{\lambda}) = 0.$$

The solutions of this equation are written conventionally as $\lambda_n^{(0)}$ $(n = 1, 2, \ldots)$. Thus the full solution is

$$u(r, t) = \sum_{n=1}^{\infty} A_n J_0(\sqrt{\lambda_n^{(0)}} r) e^{-\lambda_n^{(0)} kt}.$$

The constants A_n are determined by the initial condition $u(r, 0) = f(r)$, i.e.

$$f(r) = \sum_{n=1}^{\infty} A_n J_0(\sqrt{\lambda_n^{(0)}} r).$$

Hence A_n are the coefficients in the Fourier–Bessel series for $f(r)$.

SAQ 13

Solve the heat conduction problem in a cylindrical bar of unit radius with the initial temperature distribution given by

$$f(r) = 1 - r^2.$$

(You will need the result of SAQ 7.)

(Solution on p. 25.)

14.3 SUMMARY

This unit introduces the *Bessel function of the first kind* J_m and some of its properties. The Bessel function is defined as (a power series which is) the bounded solution of *Bessel's equation of order* m

$$x^2 \frac{d^2u}{dx^2} + x \frac{du}{dx} + (x^2 - m^2)u = 0.$$

Bessel's differential equation arises in certain eigenvalue problems and

$$\{J_m(\sqrt{\lambda_n^{(m)}}x) : n = 1, 2, \ldots\}$$

forms a (complete) set of eigenfunctions with weight function x on $[0, 1]$. Here $\lambda_n^{(m)}$ are the positive solutions of

$$J_m(\sqrt{\lambda}) = 0.$$

The Fourier series expansion of an arbitrary function in terms of these eigenfunctions is known as a *Fourier–Bessel series.*

It was shown how Bessel's equation arises when separation of variables is applied to the membrane (wave) equation and the heat conduction equation in polar co-ordinates.

Incidentally we also included elementary properties of the *Gamma function* Γ, and we related the Bessel functions of half-integral order to the elementary trigonometric functions.

14.4 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

Either put $m = 0$ in Equation (40.5) (*W: page 179*); or you can differentiate the series expansion for $J_0(t)$ term by term.

Solution to SAQ 2

We expand the derivatives on the left-hand side of Equations (40.5) and (40.6) in *W*, so that

$$-nt^{-n-1}J_n(t) + t^{-n}J_n'(t) = -t^{-n}J_{n+1}(t),$$

$$nt^{n-1}J_n(t) + t^nJ_n'(t) = t^nJ_{n-1}(t).$$

Adding t^n times the first equation to t^{-n} times the second, we obtain

$$2J_n'(t) = J_{n-1}(t) - J_{n+1}(t),$$

whence the result.

Solution to SAQ 3

The Gamma function is defined by

$$\Gamma(m) = \int_0^\infty e^{-t}t^{m-1}\,dt \qquad \text{for } m > 0.$$

(a) $\Gamma(1) = \displaystyle\int_0^\infty e^{-t}\,dt = \left[-e^{-t}\right]_0^\infty = 1$:

integration by parts gives

$$\Gamma(2) = \int_0^\infty te^{-t}\,dt = \left[-te^{-t}\right]_0^\infty + \int_0^\infty e^{-t}\,dt$$

$$= 0 + \Gamma(1) = 1;$$

and similarly

$$\Gamma(3) = \int_0^\infty t^2e^{-t}\,dt = \left[-t^2e^{-t}\right]_0^\infty + 2\int_0^\infty te^{-t}\,dt$$

$$= 0 + 2\Gamma(2) = 2.$$

(b) $\Gamma(m+1) = \displaystyle\int_0^\infty e^{-t}t^m\,dt$

$$= \left[-e^{-t}t^m\right]_0^\infty + m\int_0^\infty e^{-t}t^{m-1}\,dt$$

$$= m\Gamma(m).$$

(Note that if $m > 0$ then $\lim e^{-t}t^m = 0$ as $t \rightsquigarrow \infty$, using a result proved in Section 29.1.3 of *Unit M201 29, Laplace Transforms.*)

If m is a positive integer,

$$\Gamma(m+1) = m\Gamma(m) = m(m-1)\Gamma(m-1)$$

$$= m(m-1)\ldots 2\cdot 1\Gamma(1) \qquad \text{by induction}$$

$$= m(m-1)\ldots 2\cdot 1 = m!,$$

since $\Gamma(1) = 1$ from (a).

Solution to SAQ 4

From the definition of the Gamma function,

$$\Gamma(\tfrac{1}{2}) = \int_0^\infty e^{-t}t^{-\frac{1}{2}}\,dt.$$

Substituting $t = u^2$, we obtain

$$\Gamma(\tfrac{1}{2}) = \int_0^\infty e^{-u^2} u^{-1} 2u\, du = 2\int_0^\infty e^{-u^2}\, du$$
$$= 2 \cdot \tfrac{1}{2}\pi^{\frac{1}{2}} = \pi^{\frac{1}{2}}.$$

Solution to SAQ 5

From the series (40.4) in *W*, we have

$$t^{-m} J_m(t) = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{2^{m+2k} k!(m+k)!}.$$

Differentiating with respect to t, we find that

$$\frac{d}{dt}[t^{-m} J_m(t)] = \sum_{k=1}^{\infty} \frac{2k(-1)^k t^{2k-1}}{2^{m+2k} k!(m+k)!}$$
$$= t^{-m} \sum_{k=1}^{\infty} \frac{(-1)^k t^{m-1+2k}}{2^{m+2k-1}(k-1)!(m+k)!}$$
$$= -t^{-m} \sum_{n=0}^{\infty} \frac{(-1)^n t^{m+1+2n}}{2^{m+1+2n} n!(m+1+n)!} \qquad \text{putting } k = n+1$$
$$= -t^{-m} J_{m+1}(t),$$

which is Equation (40.5).

To obtain (40.6), we multiply the series (40.4) by t^m, obtaining

$$t^m J_m(t) = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2m+2k}}{2^{m+2k} k!(m+k)!}.$$

We differentiate both sides with respect to t, so that

$$\frac{d}{dt}[t^m J_m(t)] = \sum_{k=0}^{\infty} \frac{(-1)^k 2(m+k) t^{2m+2k-1}}{2^{m+2k} k!(m+k)!}$$
$$= t^m \sum_{k=0}^{\infty} \frac{(-1)^k t^{m-1+2k}}{2^{m-1+2k} k!(m-1+k)!}$$
$$= t^m J_{m-1}(t).$$

Solution to SAQ 6

Let

$$u(t) = \frac{1}{\pi}\int_0^\pi \cos(t \sin\theta)\, d\theta.$$

Then

$$t\frac{du}{dt}(t) = -\frac{t}{\pi}\int_0^\pi \sin\theta \sin(t\sin\theta)\, d\theta,$$

and

$$\frac{d}{dt}\left(t\frac{du}{dt}(t)\right) = -\frac{1}{\pi}\int_0^\pi \sin\theta \sin(t\sin\theta)\, d\theta - \frac{t}{\pi}\int_0^\pi \sin^2\theta \cos(t\sin\theta)\, d\theta.$$

Therefore

$$\frac{d}{dt}\left(t\frac{du}{dt}(t)\right) + tu(t) = \frac{t}{\pi}\int_0^\pi \cos^2\theta \cos(t\sin\theta)\, d\theta - \frac{1}{\pi}\int_0^\pi \sin\theta \sin(t\sin\theta)\, d\theta.$$

Now

$$\int_0^\pi \sin\theta \sin(t\sin\theta)\, d\theta = \Big[-\cos\theta \sin(t\sin\theta)\Big]_0^\pi + t\int_0^\pi \cos\theta \cos(t\sin\theta)\cos\theta\, d\theta$$
$$= t\int_0^\pi \cos^2\theta \cos(t\sin\theta)\, d\theta.$$

Hence u satisfies Bessel's equation of order zero. Clearly u is bounded as $t \rightsquigarrow 0$. Therefore

$$u(t) = AJ_0(t).$$

Now

$$u(0) = \frac{1}{\pi}\int_0^{\pi} d\theta = 1 = J_0(0),$$

so that $A = 1$ and $u(t) = J_0(t)$, as required.

Solution to SAQ 7

We require the coefficients in the series

$$1 - x^2 = \sum_{k=1}^{\infty} A_k J_0(\sqrt{\lambda_k^{(0)}}x) \qquad 0 < x < 1.$$

We multiply both sides by $xJ_0(\sqrt{\lambda_n^{(0)}}x)$ and integrate over (0, 1). By the orthogonality property of the eigenfunctions

$$A_n\int_0^1 x[J_0(\sqrt{\lambda_n^{(0)}}x)]^2\,dx = \int_0^1 (1 - x^2)xJ_0(\sqrt{\lambda_n^{(0)}}x)\,dx,$$

that is

$$\tfrac{1}{2}A_n[J_1(\sqrt{\lambda_n^{(0)}})]^2 = \int_0^1 (1 - x^2)xJ_0(\sqrt{\lambda_n^{(0)}}x)\,dx$$

using *W: page 181, line 17*. On the right-hand side the first integral has been worked out in the example in Section 14.1.2:

$$\int_0^1 xJ_0(\sqrt{\lambda_n^{(0)}}x)\,dx = \frac{1}{\sqrt{\lambda_n^{(0)}}}J_1(\sqrt{\lambda_n^{(0)}}).$$

For the other part, we make the substitution $t = \sqrt{\lambda_n^{(0)}}x$ and obtain

$$\begin{aligned}
-\int_0^1 x^3 J_0(\sqrt{\lambda_n^{(0)}}x)\,dx &= -\frac{1}{[\lambda_n^{(0)}]^2}\int_0^{\sqrt{\lambda_n^{(0)}}} t^3 J_0(t)\,dt \\
&= -\frac{1}{[\lambda_n^{(0)}]^2}\int_0^{\sqrt{\lambda_n^{(0)}}} t^2\frac{d}{dt}[tJ_1(t)]\,dt \qquad *\\
&= -\frac{1}{[\lambda_n^{(0)}]^2}\left\{\Big[t^3 J_1(t)\Big]_0^{\sqrt{\lambda_n^{(0)}}} - 2\int_0^{\sqrt{\lambda_n^{(0)}}} t^2 J_1(t)\,dt\right\} \\
&\qquad\qquad \text{integrating by parts} \\
&= -\frac{1}{[\lambda_n^{(0)}]^2}\left\{[\lambda_n^{(0)}]^{3/2}J_1(\sqrt{\lambda_n^{(0)}}) - 2\int_0^{\sqrt{\lambda_n^{(0)}}} t^2 J_1(t)\,dt\right\} \\
&= -\frac{1}{[\lambda_n^{(0)}]^2}\{[\lambda_n^{(0)}]^{3/2}J_1(\sqrt{\lambda_n^{(0)}}) - 2\lambda_n^{(0)}J_2(\sqrt{\lambda_n^{(0)}})\}. \qquad *
\end{aligned}$$

We have used *W: page 180, Equation (40.6)* in obtaining the lines marked with an asterisk. We now have

$$A_n = 4\frac{J_2(\sqrt{\lambda_n^{(0)}})}{\lambda_n^{(0)}[J_1(\sqrt{\lambda_n^{(0)}})]^2},$$

and the solution given in *W: page 422* follows.

Solution to SAQ 8

We require the coefficients A_k in the series

$$x^{m+2} = \sum_{k=1}^{\infty} A_k J_m(\sqrt{\lambda_k^{(m)}}x).$$

Multiplying both sides by $xJ_m(\sqrt{\lambda_n^{(m)}}x)$ and integrating over (0, 1), we have

$$A_n \int_0^1 x[J_m(\sqrt{\lambda_n^{(m)}}x)]^2\,dx = \int_0^1 x^{m+3}J_m(\sqrt{\lambda_n^{(m)}}x)\,dx.$$

Using *W: page 181, line 17* we find that

$$\begin{aligned}
\tfrac{1}{2}A_n[J_{m+1}(\sqrt{\lambda_n^{(m)}})]^2 &= \int_0^1 x^{m+3}J_m(\sqrt{\lambda_n^{(m)}}x)\,dx \\
&= \frac{1}{[\lambda_n^{(m)}]^{\frac{1}{2}m+2}} \int_0^{\sqrt{\lambda_n^{(m)}}} t^{m+3}J_m(t)\,dt \\
&= \frac{1}{[\lambda_n^{(m)}]^{\frac{1}{2}m+2}} \int_0^{\sqrt{\lambda_n^{(m)}}} t^2 \frac{d}{dt}[t^{m+1}J_{m+1}(t)]\,dt \qquad \text{by (40.6)} \\
&= \frac{1}{\sqrt{\lambda_n^{(m)}}}J_{m+1}(\sqrt{\lambda_n^{(m)}}) - \frac{2}{[\lambda_n^{(m)}]^{\frac{1}{2}m+2}} \int_0^{\sqrt{\lambda_n^{(m)}}} t^{m+2}J_{m+1}(t)\,dt \\
&= \frac{1}{\sqrt{\lambda_n^{(m)}}}J_{m+1}(\sqrt{\lambda_n^{(m)}}) - \frac{2}{\lambda_n^{(m)}}J_{m+2}(\sqrt{\lambda_n^{(m)}}) \qquad \text{by (40.6).}
\end{aligned}$$

Substitution of A_n back into the series gives the solution in *W: page 422.*

Solution to SAQ 9

(a) Equation (40.6) gives us

$$t^{\frac{1}{2}}J_{-\frac{1}{2}}(t) = \frac{d}{dt}[t^{\frac{1}{2}}J_{\frac{1}{2}}(t)].$$

(This expression is valid since $J_{-\frac{1}{2}}$ is defined by the series for J_m and Equation (40.6) was proved for all m in SAQ 5.) Thus

$$\begin{aligned}
t^{\frac{1}{2}}J_{-\frac{1}{2}}(t) &= \frac{d}{dt}\left[\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\sin t\right] \\
&= \left(\frac{2}{\pi}\right)^{\frac{1}{2}}\cos t.
\end{aligned}$$

Hence

$$J_{-\frac{1}{2}}(t) = \left(\frac{2}{\pi t}\right)^{\frac{1}{2}}\cos t.$$

Alternatively, if you are feeling very energetic, you could show that

$$k!\Gamma(k+\tfrac{1}{2}) = \frac{(2k)!}{2^{2k}}\pi^{\frac{1}{2}}$$

and hence derive the result by comparing power series.

(b) Equation (40.6) may be expressed as

$$t^{m-1}J_{m-1}(t) = \left(\frac{1}{t}\frac{d}{dt}\right)[t^mJ_m(t)].$$

Applying this formula n times we obtain

$$t^{m-n}J_{m-n}(t) = \left(\frac{1}{t}\frac{d}{dt}\right)^n[t^mJ_m(t)].$$

For $m = -\frac{1}{2}$ this gives us

$$\begin{aligned}
J_{-n-\frac{1}{2}}(t) &= t^{n+\frac{1}{2}}\left(\frac{1}{t}\frac{d}{dt}\right)^n[t^{-\frac{1}{2}}J_{-\frac{1}{2}}(t)] \\
&= t^n\left(\frac{2t}{\pi}\right)^{\frac{1}{2}}\left(\frac{1}{t}\frac{d}{dt}\right)^n\left(\frac{\cos t}{t}\right).
\end{aligned}$$

Solution to SAQ 10

We attempt a separable solution

$$u(r, \theta, t) = R(r)\Theta(\theta)T(t).$$

Substitution into the differential equation (41.1) gives

$$R\Theta T'' - c^2\left[R''\Theta T + \frac{1}{r}R'\Theta T + \frac{1}{r^2}R\Theta'' T\right] = 0.$$

After dividing through by $R\Theta T$ we find that

$$\frac{T''}{c^2 T} = \frac{R''}{R} + \frac{R'}{rR} + \frac{\Theta''}{r^2\Theta}.$$

Both sides must be constant, say $-\lambda$, so that

$$T'' + c^2\lambda T = 0 \tag{1}$$

and

$$\frac{R''}{R} + \frac{R'}{rR} + \frac{\Theta''}{r^2\Theta} = -\lambda.$$

The second of these equations can again be separated into

$$\frac{r^2R''}{R} + \frac{rR'}{R} + \lambda r^2 = -\frac{\Theta''}{\Theta} = m^2,$$

say. (The most convenient form of constant is usually introduced; in this case we select m^2—a square because we expect to take square roots, with a positive sign because we expect the solution for Θ to be trigonometric rather than exponential.) Thus R and Θ satisfy

$$r^2R'' + rR' + (\lambda r^2 - m^2)R = 0, \tag{2}$$

$$\Theta'' + m^2\Theta = 0. \tag{3}$$

Since the general solution of (1) is $A\cos c\sqrt{\lambda}t + B\sin c\sqrt{\lambda}t$, and

$$\frac{\partial u}{\partial t}(r, \theta, 0) = 0,$$

we choose $T(t) = \cos c\sqrt{\lambda}t$. The general solution of (3) is

$$\Theta_0(\theta) = A_0 + B_0\theta \qquad m = 0,$$

$$\Theta_m(\theta) = A_m\cos m\theta + B_m\sin m\theta \qquad m \neq 0,$$

and since $\Theta(\theta)$ must be periodic with period 2π (otherwise it would not be single-valued), m can take only the values $0, 1, 2, \ldots$ and $B_0 = 0$. Finally the solution of (2) which vanishes for $r = 1$ and is bounded at the origin is

$$J_m(\sqrt{\lambda}r),$$

the eigenvalues λ being given by the solutions of the equation

$$J_m(\sqrt{\lambda}) = 0.$$

These values are denoted by $\lambda_k^{(m)}$. Note that this initial-boundary value problem generates a doubly infinite set of eigenvalues. The series (41.3) can now be put together.

Solution to SAQ 11

The required solution is given by *W: page 182, Equation (41.3)*. Since

$$\int_{-\pi}^{\pi}\cos\theta\sin m\theta\, d\theta = 0 \qquad m = 1, 2, \ldots$$

and

$$\int_{-\pi}^{\pi}\cos\theta\cos m\theta\, d\theta = 0 \qquad m = 0, 2, 3, \ldots,$$

all the coefficients c_{km}, d_{km} vanish except c_{k1}. Thus the solution simplifies to

$$u(r, 0, t) = \sum_{k=1}^{\infty} c_{k1} \cos 0 \, J_1(\sqrt{\lambda_k^{(1)}} r) \cos(c\sqrt{\lambda_k^{(1)}} t).$$

However $f(r, 0)$ contains an *eigenfunction* for the problem, namely $J_1(\sqrt{\lambda_1^{(1)}} r)$. Consequently all the c_{k1} must vanish except c_{11}. Thus the solution reduces further to

$$u(r, 0, t) = c_{11} \cos 0 \, J_1(\sqrt{\lambda_1^{(1)}} r) \cos(c\sqrt{\lambda_1^{(1)}} t).$$

Finally, by inspection we can see that $c_{11} = 1$ in order that the initial condition should be satisfied.

Solution to SAQ 12

This is the problem of *W: page 182* with $f(r, 0) = \varepsilon(1 - r^2)$. The solution may therefore be written in the form (41.3). We require the integrals

$$\int_{-\pi}^{\pi} d\theta = 2\pi,$$

$$\int_{-\pi}^{\pi} \cos m\theta \, d\theta = 0 \qquad \text{for } m \neq 0$$

and

$$\int_{-\pi}^{\pi} \sin m\theta \, d\theta = 0.$$

Thus the solution may be expressed as

$$u(r, 0, t) = \sum_{k=1}^{\infty} A_k J_0(\sqrt{\lambda_k^{(0)}} r) \cos(c\sqrt{\lambda_k^{(0)}} t).$$

The initial condition gives

$$\varepsilon(1 - r^2) = \sum_{k=1}^{\infty} A_k J_0(\sqrt{\lambda_k^{(0)}} r);$$

using the result of SAQ 7 we find that

$$u(r, 0, t) = 4\varepsilon \sum_{k=1}^{\infty} \frac{J_2(\sqrt{\lambda_k^{(0)}}) J_0(\sqrt{\lambda_k^{(0)}} r)}{\lambda_k^{(0)} [J_1(\sqrt{\lambda_k^{(0)}})]^2} \cos(c\sqrt{\lambda_k^{(0)}} t).$$

Solution to SAQ 13

The solution is

$$u(r, t) = \sum_{n=1}^{\infty} A_n J_0(\sqrt{\lambda_n^{(0)}} r) e^{-\lambda_n^{(0)} kt},$$

with initial value

$$1 - r^2 = u(r, 0) = \sum_{n=1}^{\infty} A_n J_0(\sqrt{\lambda_n^{(0)}} r).$$

This is exactly the same problem as SAQ 7 and we can write down the solution as

$$u(r, t) = 4 \sum_{n=1}^{\infty} \frac{J_2(\sqrt{\lambda_n^{(0)}}) J_0(\sqrt{\lambda_n^{(0)}} r) e^{-\lambda_n^{(0)} kt}}{\lambda_n^{(0)} [J_1(\sqrt{\lambda_n^{(0)}})]^2}.$$

14.5 APPENDIX

Equations of Motion for Fluid Flow

In Section 1.1.3 of *Unit 1, The Wave Equation*, we derived the equation of motion for a compressible fluid moving in one dimension. We now outline briefly the derivation of the equations of motion for motion in two or three dimensions.

First we consider the **derivative following the motion.** Let the scalar field f represent some property of the fluid: we require the rate of change of $f(\mathbf{r}, t)$ at a *given fluid element*, whose position changes with time. For a general variation in $\mathbf{r} = (x, y, z)$ and t we have the *first-order Taylor approximation*

$$\Delta f = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z + \frac{\partial f}{\partial t}\Delta t$$

$$= \Delta\mathbf{r} \cdot \mathbf{grad} f + \frac{\partial f}{\partial t}\Delta t:$$

thus

$$\frac{\Delta f}{\Delta t} = \frac{\Delta \mathbf{r}}{\Delta t} \cdot \mathbf{grad} f + \frac{\partial f}{\partial t}.$$

Since we are following the motion we have

$$\lim_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \mathbf{v}(\mathbf{r}, t),$$

where $\mathbf{v}$ represents the fluid velocity. Hence, proceeding to the limit we obtain the total derivative following the motion

$$\frac{df}{dt} = \mathbf{v} \cdot \mathbf{grad} f + \frac{\partial f}{\partial t},$$

which is the three-dimensional analogue of the result obtained in *Unit 1*.

Next, we consider the generalization of the *mass-conservation equation*. (This is also known as the *equation of continuity*.) Let V denote the volume contained within a simple closed surface S. If ΔS is a small element of this surface then the volume and mass of fluid crossing ΔS per unit time are respectively $v\cos\theta\,\Delta S$ and $\rho v\cos\theta\,\Delta S$, where θ is the angle which the direction of v makes with the normal to ΔS and ρ denotes the density of the fluid. If we write $\mathbf{n}$ to denote the unit vector whose direction is that of the normal to ΔS, then the rate of flow of mass across ΔS in the sense of $\mathbf{n}$ is just $\rho\mathbf{v} \cdot \mathbf{n}\,\Delta S$. Since we adopt the convention in the case of a closed surface S that $\mathbf{n}$ is always drawn outwards, we find that

$$\int_S \rho\mathbf{v} \cdot \mathbf{n} dS = \text{mass of fluid flowing out of } S \text{ per unit time}$$

$$= \text{decrease of mass within } S \text{ per unit time}$$

$$= -\frac{\partial}{\partial t}\int_V \rho dV.$$

Now, from the Divergence Theorem (*Unit 3, Elliptical and Parabolic Equations*),

$$\int_S \rho\mathbf{v} \cdot \mathbf{n} dS = \int_V \operatorname{div}(\rho\mathbf{v})\, dV.$$

Consequently, for any volume V,

$$\int_V \operatorname{div}(\rho\mathbf{v})\, dV + \int_V \frac{\partial \rho}{\partial t} dV = 0.$$

Thus at any point of the fluid we must have

$$\operatorname{div}(\rho \mathbf{v}) + \frac{\partial \rho}{\partial t} = 0,$$

as quoted in Section 14.2.2. Since

$$\operatorname{div}(\rho \mathbf{v}) = \rho \operatorname{div} \mathbf{v} + \mathbf{v} \cdot \mathbf{grad}\, \rho,$$

we can write the expression above in terms of the total derivative as

$$\frac{d\rho}{dt} + \rho \operatorname{div} \mathbf{v} = 0.$$

Finally we consider Newton's Second Law of Motion for the fluid. Consider a small rectangular element of non-viscous fluid with sides of lengths Δx, Δy, Δz. The forces acting on it are the pressures on its several faces and possibly a body force $\mathbf{F}(\mathbf{r}, t)$ per unit volume due to gravitation or some other external force. The acceleration produced is $d\mathbf{v}/dt$ where $\mathbf{v}(\mathbf{r}, t)$ is the velocity vector $\mathbf{r}$ at some point in its interior. The x-component of the equation of motion is therefore

$$\mathbf{F}_x \Delta x \Delta y \Delta z + p \Delta y \Delta z - \left(p + \frac{\partial p}{\partial x} \Delta x\right) \Delta y \Delta z = \frac{dv_x}{dt} \rho \Delta x \Delta y \Delta z,$$

$$+ \text{ second order terms,}$$

where p denotes the pressure. Dividing by $\Delta x \Delta y \Delta z$ and letting Δx, Δy, Δz all tend to zero, we have

$$\mathbf{F}_x - \frac{\partial p}{\partial x} = \rho \frac{dv_x}{dt}$$

Combining this with the corresponding equations for y and z we have the vector equation

$$\mathbf{F} - \mathbf{grad}\, p = \rho \frac{d\mathbf{v}}{dt}.$$

In this last equation $d\mathbf{v}/dt$ is simply the vector

$$\left(\frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_x}{dt}\right),$$

and

$$\frac{dv_x}{dt} = \mathbf{v} \cdot \mathbf{grad}\, v_x + \frac{\partial v_x}{\partial t}.$$

We may condense this as

$$\frac{d\mathbf{v}}{dt} = (\mathbf{v} \cdot \mathbf{grad})\, \mathbf{v} + \frac{\partial \mathbf{v}}{\partial t},$$

so that the equation of motion becomes

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \mathbf{grad})\, \mathbf{v}\right) = -\mathbf{grad}\, p + \mathbf{F}.$$

PARTIAL DIFFERENTIAL EQUATIONS OF APPLIED MATHEMATICS

1	*W*	The Wave Equation
2	*W*	Classification and Characteristics
3	*W*	Elliptic and Parabolic Equations
4		NO TEXT
5	*S*	Finite-Difference Methods I: Initial Value Problems
6	*W*	Fourier Series
7	*N*	Motion of Overhead Electric Train Wires
8	*S*	Finite-Difference Methods II: Stability
9	*W*	Green's Functions I: Ordinary Differential Equations
10	*W*	Green's Functions II: Partial Differential Equations
11	*S*	Finite-Difference Methods III: Boundary Value Problems
12		NO TEXT
13	*W*	Sturm-Liouville Theory
14	*W*	Bessel Functions
15	*N*	Finite-Difference Methods IV: Parabolic Equations
16	*N*	Blood Flow in Arteries

The letter after the unit number indicates the relevant set book; N indicates a unit not based on either book.

Course Team

Chairman:	Professor R. C. Smith	Professor of Mathematics
Members:	Dr. A. Crilly	B.B.C.
	Mr. D. W. Jordan	University of Keele
	Dr. A. D. Lunn	Lecturer in Mathematics
	Dr. N. P. Mett	Lecturer in Mathematics
	Dr. A. G. Moss	Lecturer in Educational Technology
	Dr. D. Richards	Lecturer in Mathematics
	Mr. M. G. T. Simpson	Course Assistant
	Dr. P. Smith	University of Keele
	Dr. P. G. Thomas	Lecturer in Mathematics
	Dr. R. V. M. Zahar	Senior Lecturer in Mathematics

With assistance from:

Professor L. Fox	Oxford University
Dr. M. W. Green	University of Dundee
Professor A. Jeffrey	University of Newcastle-upon-Tyne
Mr. J. E. Phythian	Staff Tutor in Mathematics
Mr. G. D. Smith	Brunel University
Dr. T. B. Smith	Lecturer in Physics
Mr. G. Young	Staff Tutor in Mathematics